Analyse de données transcriptome et protéome pour l'étude des réponses aux stress oxydants et aux métaux lourds

Magali Michaut

N° d'ordre : 9268

École doctorale d'Informatique
de Paris-Sud

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS XI

Spécialité Informatique

À présenter et soutenir publiquement par

Magali MICHAUT
Le 28 novembre 2008

Analyse de données transcriptome et protéome pour l'étude des réponses aux stress oxydants et aux métaux lourds

Directeur de thèse
Dr Pierre LEGRAIN

JURY

Rapporteurs :	Pr Florence d'ALCHÉ-BUC
	Dr Christine BRUN
	Dr Daniel KAHN
Examinateurs :	Pr Alain DENISE
	Dr Jacques VAN HELDEN
Encadrants :	Dr Jean-Christophe AUDE
	Dr Pierre LEGRAIN

Chaque rythme et chaque son étant d'une profondeur ineffable,
les mots sont inutiles à qui comprend la musique.

Parabole du bouddhisme zen.

Préface

Face à la montagne de remerciements que j'ai envie de faire, je ne sais par où commencer. J'essaye de dégager une structure logique pour présenter les choses : géographique ? alphabétique ? chronologique ? aléatoire ?... déformation professionnelle ! C'est vrai qu'il s'en est passé des choses ces dernières années.

Tout commence un jour de l'année 2003, alors que je suis en première année d'école d'ingénieurs à Supélec. À la fin du dernier cours d'ouverture de "Génétique et biologie moléculaire", je me décide enfin à poser à l'intervenant la question qui me trotte dans la tête depuis quelques temps : "J'aimerais faire de la bioinformatique après Supélec, comment je dois faire ?". Michel Vervoot, que je remercie beaucoup, m'indique deux personnes à aller voir. C'est là que je rencontre Laure Vescovo, en thèse de bioinformatique dans les locaux de Supélec, qui me renseigne très gentiment et me passe un bouquin bien intéressant. Je vais ensuite jusqu'au campus de la fac d'Orsay pour discuter avec Olivier Lespinet, que j'ai maintenant l'occasion de croiser de temps en temps, et que je remercie chaleureusement.

Un an a passé, j'en connais un peu plus sur les transistors C-MOS, mais pas tellement sur la bioinformatique... Heureusement, je profite de mon stage de fin de deuxième année pour découvrir un nouveau territoire : le Commissariat à l'Énergie Atomique (CEA pour les intimes). Je passe l'été 2004 devant un ordi à découper des hiérarchies pour classer des gènes (ou des choux-fleurs, ça marcherait aussi, à condition de définir la distance entre deux choux-fleurs). Merci à Jean-Christophe de m'avoir accueillie en stage. Merci à Laure de m'avoir accompagnée dans cette première expérience. Merci pour ta bonne humeur, pour les conseils, et aussi pour les parties de tennis malheureusement trop peu nombreuses. Suite à cette expérience estivale enrichissante, j'ai bien envie de revenir. Et me voilà, quelques mois plus tard, en train de faire mon stage de master au Laboratoire de Physiogénomique du Service de Biochimie et Génétique Moléculaire du Département de Biologie Joliot-Curie de la Direction des Sciences du Vivant ... ouf ! En s'entraînant, on peut même le dire dans l'autre sens et très vite. "T'es quoi, toi ? Ben moi j'suis DSV/DBJC/SBGM/LPG"... et toc ! J'aimerais remercier Carine Audubert, qui était en stage dans le labo à ce moment et m'a fait découvrir les parsers BioWarehouse. Merci pour ta bonne humeur et ta gentillesse.

L'été passe et me voici prête pour commencer ma thèse avec Jean-Christophe et sous la direction d'un certain Pierre Legrain. Octobre 2005, le projet de thèse est mis en place au sein d'une collaboration avec Franck Chauvat et Corinne Cassier-Chauvat, du Service de Biologie Moléculaire et Systémique. Je suis financée par un Contrat de Formation par la Recherche grâce au CEA et à l'INSTN (Institut National des Sciences et Techniques Nucléaires). Je tiens à remercier Pierre Legrain, à l'époque chef du DBJC, et Michel Werner, à l'époque chef du SBGM, de m'avoir accueillie dans le service pour faire ma thèse. Je voudrais également remercier Franck et Corinne d'avoir participé à ce projet

et de m'avoir aidée à apprendre plein de choses en biologie. Je tiens à remercier aussi Laetitia Houot, avec qui j'ai parlé puces (à ADN j'entends), et que j'ai eu le plaisir de revoir à Boston quelques années plus tard. Merci aussi aux autres "cyano" qui ont suivi : Martial Marbouty, Kinsley Narainsamy, Sandrine Farci, Cyril Saguez, Arounie Tavenet.

Assez vite, le bureau se remplit. Olivier Delalande fait semblant de faire des calculs très longs sur l'ordi d'à côté. Merci à toi pour les belles images de protéines, et surtout pour l'histoire du petit Xénon qui se retrouve piégé dans la protéine... mémorable ! Bien sûr, il y a les petites visites du père Boulard, Yves de son prénom. Toujours des fausses rumeurs à colporter, des prétextes pour râler, mais au fond, on sait bien que c'est pour prendre gentiment des nouvelles et dire bonjour. D'ailleurs, Yves, le tennis, c'est quand tu veux ! Et puis il y a Guillaume Meurice, soutien formidable, au cours de discussions scientifiques et au-delà : les Fatals Picards, le badminton, le Fond Qui Cache, j'en passe et des meilleures. Un grand merci à toi.

Petit à petit, mes interactions s'étendent même en-dehors du fameux "couloir de la mort" du bâtiment 144 (là où y'a des gens bizarre qui restent devant un ordi toute la journée et où personne ne vient puisqu'il se situe après la cafet...). Merci en particulier à Yad Ghavi-Helm, avec qui c'était un plaisir de collaborer, et qui m'a appris à utiliser une pipette, faire un gel, un filtrage sur colonne. J'ai même fait une quantification d'ADN par spectrophotométrie... j'suis trop fière ! Bon d'accord, il suffit d'appuyer sur un bouton, mais le nom ça fait sérieux. Et puis merci aussi à Audrey Suleau, Olivier Lefebvre, Christine Conesa, Joël Acker pour le projet intéressant sur lequel j'ai pu travailler. Merci également à Christian Marck pour les déjeuners instructifs et les histoires toujours passionnantes. Merci à Pierre Thuriaux pour l'ADN de testicules de saumon, que j'ai pu montrer aux jeunes collégiens d'Evry. "Ouaich... on dirait un vieux bout de Kleenex !". Merci encore à Julie Soutourina, Noëlle Dufour, Etienne Thévenot, Anne Peyroche, Benoît Le Tallec, Marie-Claude Marsolier-Kergoat, Christophe Carles, Michel Riva, Nayla Ayoub et Carl Mann, que j'ai côtoyés dans le service.

Avec l'été 2006 arrivent les stagiaires et la canicule. Merci à Manuel Montoya-Collin, Emmanuelle Billon, Marion Verdenaud. Vient alors le moment de faire la fête au DBJC, non pardon, d'organiser les journées des thésards, et de faire un petit concert. Un grand merci à Véronique Berthonaud, Daphné Despres et Pierre Legrain, pour le partage musical, et à Élise Pouchelet, qui nous a fait bien progresser. Dès la rentrée, les grands projets reprennent ; feu-le-DBJC fait maintenant partie du nouvel Institut de Biologie et de Technologies de Saclay (iBiTec-S) ; le SBGM gagne deux voyelles et devient le SBIGeM (Service de Biologie Intégrative et Génétique Moléculaire) ; nous faisons alors partie du Laboratoire de Biologie Intégrative (LBI) et l'équipe des BNs, non DBN, se crée : Dynamics of Biological Networks, et ouais, nous on parle anglais, ça fait classe ! Merci à Jean Labarre, Peggy Baudouin-Cornu (avec un u parce que sinon je vais me faire gronder), Stéphane Chédin, Gilles Lagniel, Manu Godat et Stéphanie Boisnard, ainsi qu'à Jean-Yves Thuret et Régis Courbeyrette pour les repas animés et les bons conseils.

L'année 2007 est pour moi synonyme de nouveaux horizons. Dès le 2 janvier (sic !), je m'envole pour Cambridge en Angleterre. J'aimerais adresser un immense merci à Rolf Apweiler et Henning Hermjakob de m'avoir accueillie à l'Institut Européen de Bioinformatique (EBI) à Hinxton. J'ai trouvé très agréable et surtout très stimulant de travailler au sein de l'équipe IntAct. J'aimerais remercier tout particulièrement Samuel Kerrien, qui m'a soutenue pendant tout mon séjour, sur les plans professionnel et personnel, Catherine Leroy qui m'a chaleureusement accueillie, ainsi que Nicolas Rodrigues

et Phillippe Aldebert. Et puis aussi un grand merci à Luisa Montecchi-Palazzi pour l'accompagnement sur le projet, ainsi qu'à Robert Petryszak et Paul Kersey. Un merci particulier à Nils Gehlenborg qui m'a fait découvrir l'ISCB Student Council dans lequel je suis maintenant largement engagée. Un grand merci à la fameuse et brillante équipe de foot EuroTrash, Thomas Laurent, Gautier Koscielny et les autres ; Emilio Salazar pour m'avoir permis de jouer un match de foot Oxford vs Cambridge ; Vincent Le Texier pour les parties de tennis ; je n'oublie pas non plus Markus Brosch qui m'a sauvé la vie quand mon ordi à craqué ; merci aussi à Florence Cavalli, en particulier pour le Darwin May Ball, le punting devant King's College la nuit... simplement magique. Difficile de citer tous les gens que j'ai eu la chance de rencontrer sur le campus, de l'EBI ou du Sanger Institute. Merci à Bruno Aranda, David Gloriam, Michael Mueller, Florian Reisinger, Jyoti Khadake, Michael Kleen, Samuel Patient, Richard Côté, Sandra Orchard, Cathy Derow, Lennart Martens, Karyn Megy, Nicolas Le Novère, Nick Luscombe, Fabio Pardi, Wolfgang Huber, Antony Quinn, Nataliya Sklyar. Ce séjour chez les Anglais a été un vrai bonus pour la thèse.

En juin 2007, une réunion de travail stratégiquement planifiée me permet de participer au foot de l'iBiTec-S. Merci à l'équipe du SBIGeM, invaincue (imbattable ?) depuis lors. Avec les beaux jours, je repointe le bout de mon nez au CEA, comme le font les petits lapins (grâce à qui Saclay reste rose). Un grand merci à Catherine Doreau et à Chantal Le Gourrierec, pour la patience avec laquelle elles ont répondu aux innombrables questions que je suis venue leur poser au secrétariat. De retour de postdoc aux États-Unis, Agnès Delaunay-Moisan me tient compagnie (quand elle n'est pas en train de s'occuper des pattes des souris). Merci à toi pour tous les conseils et les discussions bien sympathiques. Merci aussi à Fred Tacnet, et bonne continuation à toi.

Le boulot continue et InteroPorc est mis à la disposition de tous, grâce à l'interface web développée par Arnaud Martel du GIPSI (Groupement Informatique Pour les Scientifiques d'Ile de france). Un énorme merci à toi pour ta motivation et ton efficacité qui ont très largement contribué à la valorisation de mon travail. Merci également à Etienne Formstecher qui m'a bien aidée avec les interactions d'*H. pylori*. Merci beaucoup à Raphaël Guérois pour les conseils avisés, la relecture du papier et les échanges d'expériences toujours très intéressants. Et merci à Alain Denise pour son oreille attentive depuis le master et tout au long de ma thèse. J'en profite pour remercier également les membres du jury qui ont accepté d'évaluer mon travail : Florence d'Alché-Buc, Christine Brun, Alain Denise, Daniel Kahn, Jacques Van Helden.

Plus généralement, je voudrais te remercier, Jean-Christophe, de m'avoir encadrée sur la globalité de ce projet. J'ai apprécié en particulier l'ambiance de travail, ton ouverture d'esprit et la confiance que tu m'a accordée. J'aimerais également te remercier, Pierre, pour l'attention que tu as porté à mon travail malgré tes responsabilités grandissantes. J'ai beaucoup apprécié ta manière de t'adapter à mon caractère, de me rassurer et de détecter mes haussements de sourcils peu perceptibles. Je voudrais remercier aussi toutes les personnes qui, inconsciemment ou non, ont parsemé ce chemin d'embûches, m'apprenant ainsi à accepter les contraintes de "la vraie vie"...

Finalement, l'été revient et les occupations ne manquent pas. Merci à Arthur Moisdon et Marianne Hervé, venus en stage au labo et qui ont fait du très bon boulot. Merci à la Ch'tite Equipe Atomique pour les agréables moments passés à Lille : Samia Aci, Paul Garcin, Florence Combes qui m'a fait découvrir en avant-première le tamagoshi du CEA (que certains appellent aussi MobiPass), Sylvain Bournais, Susete Alves-Carvalho, sans oublier Yves Vandenbrook et sa bouille incroyable après quelques verres. J'aimerais

aussi remercier Gary Bader, Andrea Califano, Erin O'Shea, Roy Kishony, Martha Bulyk, Manolis Kellis, Fritz Roth, Marc Vidal et Ilya Shmulevich de m'avoir invitée à donner un séminaire lors de ma "tournée des labos", à la recherche du postdoc perdu (finalement retrouvé à Toronto). Un grand merci également à Céline Lefebvre pour l'accueil et l'hébergement royal à New York. Merci beaucoup à Mike Calderwood pour l'accueil chaleureux et l'hébergement à Boston.

Arrive le moment d'écrire tout ça dans le manuscrit. Un grand merci à mon tonton Jean-Yves pour ces remarques d'expert. Merci également à ma cousine Marine, qui m'a indiqué le chemin vers le lycée Saint-Louis, puis le CEA. Merci pour tes conseils. J'aimerais remercier du fond du coeur Benjamin Duval, qui a pris la peine de lire le manuscrit en entier et qui a contribué à son amélioration. Au-delà de ça, je voudrais te remercier infiniment de me supporter au quotidien (dans les deux sens du terme), et de comprendre mon engagement dans le travail. Je finis par un bisou à ma famille nucléaire, Elouan, Agnès, Manu, Mélanie et bien sûr mon Papa, ma Maman, mon poisson rouge !

Magali Michaut, Université Paris-Sud 11, 18 novembre 2008

INTITULÉ ET ADRESSE DU LABORATOIRE

Laboratoire de Biologie Intégrative

Service de Biologie Intégrative et de Génétique Moléculaire

Institut de Biologie et de Technologies de Saclay

Bâtiment 142, Pièce 29, Point courrier n°21

CEA Saclay, 91191 Gif Sur Yvette, France

Abstract

Title : Transcriptomic and proteomic data analysis to study responses to oxidative stress and heavy metals

This work aims at studying responses to oxidative stress and heavy metals through transcriptomic and proteomic data analysis, in particular in the cyanobacterium *Synechocystis*. This organism is a prokaryote largely studied which notably enables to improve the understanding of plants and is easy to manipulate genetically.

The approach first involved analysing the transcriptional responses of *Synechocystis*' genes in stress conditions, particularly in the presence of cadmium or hydrogen peroxide. Methods to predict protein-protein interactions were then developed in order to construct an interaction network. This network was compared to an experimental network in terms of structure. It was then complemented with transcriptomic data previously analysed in order to obtain a more integrated view of the different phenomena and to study the dynamics of functional modules.

The results show different phases in the transcriptional responses as well as functional groups of interacting and co-expressed proteins. In addition, the automation of a mixed hierarchical-pyramidal classification method is proposed. A method to identify composition biases between groups of proteins was also developed. Furthermore, a protein-protein interaction prediction tool was developed, of use for all sequenced species. This open-source software, InteroPorc, has been made available and has the great advantage of being flexible since it can be applied to different source interactions. Furthermore this tool can be easily run online through a web interface (http ://biodev.extra.cea.fr/interoporc/).

Keywords : bioinformatics, integration, prediction, transcriptome, proteome, interactions, network, software

Résumé

Titre : Analyse de données transcriptome et protéome pour l'étude des réponses aux stress oxydants et aux métaux lourds

Ce travail a pour objet l'analyse de données transcriptome et protéome pour l'étude des réponses aux stress oxydants et aux métaux lourds, en particulier chez la cyanobactérie *Synechocystis*. Cet organisme procaryote permet notamment d'aider à la compréhension des plantes tout en étant facilement manipulable génétiquement.

La démarche a d'abord consisté à analyser les réponses transcriptionnelles des gènes de *Synechocystis* en conditions de stress, notamment en présence de cadmium ou de peroxyde d'hydrogène. Des méthodes de prédiction d'interactions protéine-protéine ont ensuite été développées afin de construire un réseau d'interactions. Ce dernier a été comparé à un réseau d'interactions identifiées expérimentalement, notamment en termes de structure. Puis il a été complété avec les données de transcriptome précédemment analysées, afin d'obtenir une vision plus intégrée des différents phénomènes et d'étudier la dynamique des modules fonctionnels.

Les résultats font apparaître différentes phases dans les réponses transcriptionnelles, ainsi que des groupes fonctionnels de protéines en interaction et co-exprimées. De plus, l'automatisation d'une méthode de classification mixte hiérarchique-pyramidale est proposée. Une méthode d'identification de biais de composition entre des groupes de protéines a aussi été développée. Par ailleurs, un outil de prédiction d'interactions protéine-protéine, applicable à toutes les espèces séquencées, a été développé. Ce logiciel open-source, InteroPorc, présente l'avantage d'être flexible, puisqu'il peut s'appliquer à différents jeux d'interactions sources. En outre, l'outil est facilement utilisable en ligne à travers une interface web.

Mots-clés : bioinformatique, intégration, prédiction, transcriptome, protéome, interactions, réseau, logiciel

Table des matières

Liste des tableaux

Table des figures

Terminologie

Les termes en gras sont indiqués dans cette terminologie.

ADN L'acide désoxyribonucléique, ou **ADN**, est une molécule présente dans toutes les cellules vivantes. L'**ADN** est le support de l'hérédité ou de l'information génétique car il constitue le **génome** des êtres vivants et se transmet en totalité ou en partie lors des processus de reproduction. L'**ADN** détermine la synthèse des **protéines**.

ARN L'acide ribonucléique, ou **ARN**, est une molécule similaire à l'**ADN**, aussi bien en termes structurels qu'en termes fonctionnels (matérialisation et traitement de l'information génétique). Il existe différents type d'ARN, notamment l'**ARN** messager (ARNm) qui est formé par transcription de l'**ADN** dont il est la copie. Son rôle consiste à transporter l'information génétique recueillie du noyau vers le cytoplasme, où elle sera traduite en **protéine** par les **ribosomes**.

Acide aminé Un acide aminé est une molécule organique possédant un squelette carboné et deux fonctions : une amine (-NH2) et un acide carboxylique (-COOH). Les acides aminés sont les unités structurales de base des **protéines**.

Arbre Un **arbre** est un graphe connexe sans cycle. Dans un **arbre**, on distingue deux catégories d'éléments : les **feuilles**, éléments ne possédant pas d'**enfant** dans l'**arbre** ; les nœuds internes, éléments possédant des **enfants** (sous-branches). La **racine** de l'**arbre** est le nœud ne possédant pas de **parent**.

Codon Un **codon** est un triplet de **nucléotides** A, C, U ou G de l'**ARN** messager.

Enfant Dans la théorie des graphes, et en particulier dans le cas des **arbres**, on définit une orientation (de la **racine** vers les **feuilles**, la **racine** se situant en général en haut). Pour un nœud donné, le nœud au-dessus est le **parent** ou père et les nœuds en-dessous sont les **enfants**.

Feuille Dans la théorie des graphes, et en particulier des **arbres**, une **feuille** fait référence à un nœud n'ayant aucun **enfant** (les **feuilles** se situent en général en bas).

Génome Le **génome** est l'ensemble du matériel génétique d'un individu ou d'une espèce encodé dans son **ADN** (à l'exception de certains virus dont le **génome** est porté par des molécules d'**ARN**).

Gène Un gène est une séquence d'**ADN** qui spécifie la synthèse d'une chaîne de polypeptide ou d'un **ARN** fonctionnel.

Graphe planaire Dans la théorie des graphes, un **graphe planaire** est un graphe qui peut être représenté sur un plan sans qu'aucune arête n'en croise une autre.

Graphlet Dans la théorie des graphes, un **graphlet** est un petit sous-graphe induit. Un sous-graphe induit par un ensemble de sommets est constitué de tous ces sommets et de toutes les arêtes reliant deux de ces sommets.

Homologue En biologie de l'évolution, une **homologie** désigne une similarité entre deux traits (en général anatomiques) observés chez deux espèces différentes, qui est due au fait que toutes deux l'ont hérité d'un ancêtre commun. Ces traits sont alors dits **homologues**. Ce peut être des caractères anatomiques ou moléculaires (**protéines homologues**). Ce terme s'étend aussi aux séquences d'**ADN**.

Interactome Le terme **interactome** fait référence à l'ensemble des interactions pouvant avoir lieu entre des **protéines** d'une cellule donnée.

Nucléotide Les **nucléotides** sont des acides désoxyribonucléiques pour l'**ADN** (A, C, G, T) et ribonucléiques (A, C, G, U) pour l'**ARN**.

Orthologue En biologie de l'évolution, deux **protéines** sont dites **orthologues** lorsqu'elles dérivent d'un ancêtre commun après un événement de spéciation.

Paralogue En biologie de l'évolution, deux **protéines** sont dites **paralogues** lorsqu'elles dérivent d'un ancêtre commun après un événement de duplication.

Parent Voir **Enfant**

Protéine Une **protéine** est une macromolécule composée par une ou plusieurs chaînes, ou séquences, d'**acides aminés** liés entre eux par des liaisons peptidiques.

Protéome Le **protéome** est la carte de l'ensemble des **protéines** produites par un **génome**.

Racine Dans la théorie des graphes, et en particulier des **arbres**, la **racine** fait référence à l'unique nœud n'ayant aucun **parent** (il se situe en général en haut).

Ribosome Un **ribosome** est un complexe formé de **protéines** et d'**ARNs** ribosomiques. Sa fonction est de synthétiser les **protéines** en décodant l'information contenue dans l'**ARN** messager.

Transcriptome Le **transcriptome** est l'ensemble des transcrits (**ARNs**) présents dans une cellule.

Abréviations

La liste suivante indique des abréviations qui sont utilisées dans ce manuscrit.

- ADN : Acide Désoxyribo-Nucléique
- ARN : Acide Ribo-Nucléique
- AD : domaine d'activation (*Activation Domain*)
- AP-MS : spectrométrie de masse de complexes purifiés (*Affinity Purification - Mass Spectrometry*)
- BD : domaine de liaison (*Binding Domain*)
- CAH : Classification Ascendante Hiérarchique
- CAP : Classification Ascendante Pyramidale
- CDK : kinase dépendante de la cycline (*Cyclin-Dependant Kinases*)
- COG : classe de groupes orthologues (*Cluster of Orthologous Groups*)
- DIWV : indice d'instabilité des dipeptides (*Dipeptide Instability Weight Value*)
- DDI : interaction domaine-domaine (*Domain Domain Interaction*)
- HUPO : organisation sur le protéome humain (*Human Proteome Organization*)
- IST : séquence étiquette d'une interaction (*Interaction Sequence Tag*)
- MCL : algorithme de classification de Markov (*Markov Cluster Algorithm*)
- OBO : ontologies biomédicales ouvertes (*Open Biomedical Ontologies*)
- OLS : service de consultation d'ontologie (*Ontology Lookup Service*)
- ORF : cadre ouvert de lecture (*Open Reading Frame*)
- PCR : réaction en chaîne de la polymérase (*Polymerase Chain Reaction*)
- PORC : classe d'orthologues potentiels (*Putative ORthologous Cluster*)
- PPI : interaction protéine-protéine (*Protein-Protein Interaction*)
- PSI : initiative sur les standards en protéomique (*Proteomics Standard Initiative*)
- PDB : banque de données de protéines (*Protein Data Bank*)
- ROS : espèce réactive de l'oxygène (*Reactive Oxygen Species*)
- VB : appât viable (*Viable Bait*)
- VBO : appât viable seulement (*Viable Bait Only*)
- VBP : appât/proie viable (*Viable Bait/Prey*)
- VP : proie viable (*Viable Prey*)
- VPO : proie viable seulement (*Viable Prey Only*)
- Y2H : double-hybride chez la levure (*Yeast 2-Hybrid*)

Introduction

Parfois je rêve de découvrir le secret des gènes,
mais pas une fois je n'ai eu la moindre trace d'idée respectable.

James D Watson,
The double helix, 1968

Les organismes vivants sont en interaction complexe avec leur environnement. Or, celui-ci est amené à fluctuer en permanence, ce qui peut, dans certains cas, induire des stress pour les organismes qui subissent ces variations. Lorsque les organismes sont soumis à un stress, notamment des stress oxydants ou métalliques, des modifications ont été observées à différents niveaux, en particulier physiologique, cellulaire ou métabolique. Pour comprendre les mécanismes cellulaires qui sont à la base de ces modifications, des expérimentations ont été entreprises afin d'observer les organismes soumis à un stress. Pour cela, des techniques expérimentales ont été mises en place depuis de nombreuses années comme par exemple les techniques d'hybridation d'ADN et de séparation de protéines [Klipp *et al.*, 2005].

En plus de ces techniques élémentaires qui sont toujours utilisées, d'autres techniques ont été développées plus récemment, comme par exemple l'amplification d'ADN par PCR, les puces à ADN ou à protéines, les méthodes de double-hybride, la spectrométrie de masse ou encore l'interférence ARN. La plupart de ces techniques avancées ont pu être automatisées, permettant ainsi un changement d'échelle important. En effet, lorsque les techniques élémentaires permettaient d'étudier les phénomènes à l'échelle d'un petit nombre de gènes ou de protéines, l'automatisation de ces techniques plus avancées fournit un moyen d'étude au niveau de l'ensemble des gènes d'une cellule, c'est-à-dire du génome, ou des protéines, c'est-à-dire du protéome. Par conséquent, la majorité de ces nouvelles techniques expérimentales fournissent un grand volume d'information. Pour cette raison, elles sont qualifiées de méthodes à haut-débit. Ces techniques expérimentales permettent de quantifier certaines grandeurs comme le taux d'expression des gènes (puces à ADN) ou la capacité qu'ont deux protéines à interagir (double-hybride).

Cependant, les difficultés techniques inhérentes à ces expérimentations conduisent à la production de données plus ou moins bruitées, biaisées et incomplètes. De plus, chaque technique expérimentale est développée afin d'observer un phénomène particulier, comme par exemple la régulation de la transcription ou les interactions entre protéines. Par conséquent, chacune de ces techniques permet d'obtenir des données de na-

ture différente, comme par exemple des données concernant le transcriptome, c'est-à-dire l'ensemble des transcrits d'une cellule à un moment donné, ou concernant l'interactome, c'est-à-dire l'ensemble des interactions possibles dans une cellule entre des protéines. Ces données rendent compte séparément, et à des échelles temporelle et spatiale éventuellement différentes, de plusieurs aspects des réponses cellulaires étudiées.

Ces limites conduisent à une compréhension limitée, car partielle, des réponses cellulaires étudiées. De plus, le manque de données expérimentales empêche actuellement d'étudier ces réponses chez certaines espèces, en particulier chez *Synechocystis*. Pourtant, *Synechocystis* est reconnu comme étant un organisme modèle qui présente beaucoup d'intérêts. Cette bactérie est en particulier bien adaptée à la génomique fonctionnelle car elle est dotée d'un petit génome entièrement séquencé [Kaneko *et al.*, 1996], et des outils ont été développés afin de la manipuler génétiquement [Domain *et al.*, 2004], [Mazouni *et al.*, 2004]. *Synechocystis* est d'ailleurs l'un des premiers organismes, avec la levure, dont le génome entier a été séquencé. De plus, *Synechocystis* est une cyanobactérie et possède donc un grand nombre de gènes en commun avec les plantes [Martin *et al.*, 2002]. Par conséquent, la compréhension de ses réponses aux stress oxydants et aux métaux lourds peut aider à comprendre comment les plantes réagissent face à ces stress environnementaux.

Dans ce contexte, des méthodes ont été développées pour pallier les limites liées aux techniques expérimentales. Pour cela, certains travaux se sont concentrés sur un type de données particulier. Ainsi, Eisen *et al.* ont proposé les premières approches d'analyses de données transcriptome et une manière de classer les gènes [Eisen *et al.*, 1998]. Par la suite, Quackenbush *et al.* ont développé des méthodes d'analyses de données transcriptome obtenues par puces à ADN [Quackenbush, 2001], [Quackenbush, 2002]. Ces méthodes présentent en particulier l'avantage de prendre en compte les biais et le bruit en utilisant des modèles de manière à identifier les gènes régulés. Cependant, peu de données expérimentales de transcriptome sont disponibles pour *Synechocystis*. Dans le cas des interactions protéine-protéine, pour pallier le manque de données expérimentales, Walhout *et al.* ont développé des méthodes de prédiction *in-silico* [Walhout *et al.*, 2000]. Celles-ci sont efficaces et ont permis d'étendre les réseaux d'interactions protéine-protéine de plusieurs espèces comme la levure ou le ver. Pourtant, ces méthodes de prédiction n'ont été appliquées jusqu'à présent qu'à un nombre restreint d'organismes modèles.

D'autres travaux se sont concentrés sur l'intégration de différents types de données. En particulier, Troyanskaya *et al.* ont développé des méthodes pour intégrer des données de transcriptome et d'interactome afin d'inférer des relations fonctionnelles entre les protéines [Myers et Troyanskaya, 2007]. Schramm *et al.* ont également combiné des données transcriptome et interactome afin de mieux comprendre les mécanismes d'adaptation à un changement de l'environnement [Schramm *et al.*, 2007]. Ces auteurs ont obtenu des résultats encourageants sur des organismes modèles comme la levure ou la bactérie *E. coli*. Toutefois, le manque de données évoqué précédemment chez *Synechocystis*, aussi bien au niveau du transcriptome que de l'interactome, est une limite forte pour appliquer ces méthodes d'intégration.

En définitive, l'ensemble des approches développées a permis de normaliser les données expérimentales, les compléter avec des prédictions et les combiner. Cependant, peu de données sont généralement disponibles pour *Synechocystis* jusqu'à présent, rendant cette intégration impossible.

C'est ce qui justifie ce travail. Il consiste à analyser des données transcriptome et protéome pour étudier les réponses aux stress oxydants et aux métaux lourds.

La démarche a consisté à analyser les réponses transcriptionnelles des gènes et à compléter des données d'interactions protéine-protéine pour ensuite combiner ces informations dans le but d'obtenir une vision plus complète. Dans un premier temps, l'objectif a été de décrire la régulation de la transcription chez *Synechocystis* soumis à une altération de son environnement (chapitre 1). Cette description a été effectuée en réponse à des stress oxydants et métalliques, notamment un excès de cadmium, de peroxyde d'hydrogène, de fer, de zinc, ou une carence en fer. Pour cela, les gènes induits et réprimés ont été identifiés afin de caractériser la réponse cellulaire de manière globale [Houot *et al.*, 2007]. Les principales catégories de gènes intervenant dans la réponse aux stress oxydants et métalliques ont ensuite été identifiées en considérant les classes de gènes co-exprimés. Pour cela, une méthode de classification mixte hiérarchique-pyramidale a été développée [Polaillon *et al.*, 2007]. Enfin l'idée a été d'identifier des éventuels biais de composition entre deux groupes de protéines dans le but de dégager des tendances globales qui pourraient expliquer en partie la régulation de la transcription. Au-delà des gènes régulés, nous avons ensuite voulu étudier les protéines codées par ces gènes et les interactions entre elles. En effet, les liens physiques entre les protéines apportent une information complémentaire aux liens fonctionnels entre les gènes.

La faible quantité d'interactions protéine-protéine expérimentalement identifiées chez *Synechocystis* a conduit à étudier les méthodes de prédiction d'interactions (chapitre 2). L'objectif a alors été d'adapter et de développer des méthodes de prédictions *in-silico* dans le but de construire un réseau d'interactions pour *Synechocystis* [Michaut *et al.*, 2007]. À cette occasion, un outil de prédiction a été développé et mis à la disposition de la communauté [Michaut *et al.*, 2008d]. Pendant que ce travail de prédiction était effectué [Michaut *et al.*, 2008b], des interactions expérimentales ont été publiées pour *Synechocystis* [Sato *et al.*, 2007] alors qu'aucune étude à grande échelle n'avait été réalisée pour *Synechocystis* au début du projet. Ainsi, l'idée suivante a été d'exploiter ces nouvelles données expérimentales.

L'analyse des données expérimentales de Sato *et al.* s'est faite en deux temps (chapitre 3). Dans un premier temps, les données expérimentales ont été analysées séparément afin de mettre en évidence leurs forces et leurs faiblesses, en particulier les problèmes d'asymétrie. Dans un second temps, une comparaison avec les prédictions *in-silico* a été menée selon deux étapes principales. L'organisation globale a d'abord été analysée à l'aide de critères topologiques. C'est ensuite l'organisation locale qui a été étudiée en termes de modules [Michaut *et al.*, 2008a]. Les relations entre les protéines étant jusqu'alors statiques, l'idée a ensuite été d'étudier la dynamique de ces relations.

Pour décrire la dynamique des relations entre les protéines, le réseau a été décomposé en modules fonctionnels en identifiant des classes de protéines potentiellement en

interaction et co-exprimées (chapitre 4). L'objectif a ensuite été de voir comment ces modules fonctionnels évoluent en termes de composition au cours d'une réponse à un stress oxydant ou métallique, notamment en caractérisant les transitions d'un temps à un autre d'une cinétique.

Première partie

Étude bibliographique

Chapitre 1

Analyses de données transcriptome

"Si j'ai vu plus loin que les autres,
c'est parce que j'ai été porté par des épaules de géants."

Sir Isaac Newton,
Lettre à Robert Hooke, 1676

L'objectif des expériences dites de transcriptome est principalement de mesurer l'abondance des acides ribo-nucléiques messagers (ARNm) pour un grand nombre de gènes de manière simultanée. Dans le cas d'une analyse différentielle, cette technique permet en particulier de comparer l'expression des gènes dans différentes conditions. Une condition dite standard est souvent utilisée comme contrôle afin d'étudier la condition expérimentale voulue, telle que la présence de cadmium. Ceci permet notamment d'analyser la régulation de l'expression des gènes en réponse à cette condition spécifique.

Pour cela, la technique la plus courante consiste à évaluer le niveau d'expression des gènes par les puces à ADN. Ce terme puce fait référence à une petite plaque de verre, de silicium ou encore de plastique. Une seule puce à ADN permet de refléter le niveau d'expression de milliers de gènes à un moment donné. En effet, une puce est constituée d'un grand nombre de petites zones, appelées spots, dans lesquelles sont déposés des éléments connus, en général des gènes ou des fragments de gènes. Ces puces à ADN peuvent être séparées en trois catégories selon le mode de dépôt, le nombre de conditions testées et le type de marquage utilisé : les macroarrays (dépôt direct, une condition, marquage par radioactivité), les microarrays (dépôt direct, deux conditions, marquage par fluorescence) et les puces à oligonucléotides (synthèse *in-situ*, une condition, marquage par fluorescence). Expliquons rapidement le principe général de cette technique expérimentale dans le cas d'une analyse différentielle utilisant des micrroarrays (voir Figure 1.1).

Une librairie de clones d'ADNc distincts (ADN complémentaire) est tout d'abord amplifiée par la technique de PCR, puis fixée sur la puce. Deux échantillons d'ARNm sont ensuite extraits des cellules à partir de différentes conditions. Une condition peut être spécifique par exemple d'un tissu, d'un stade de développement, d'un état de maladie, ou encore d'un traitement particulier, comme par exemple l'exposition au cadmium à

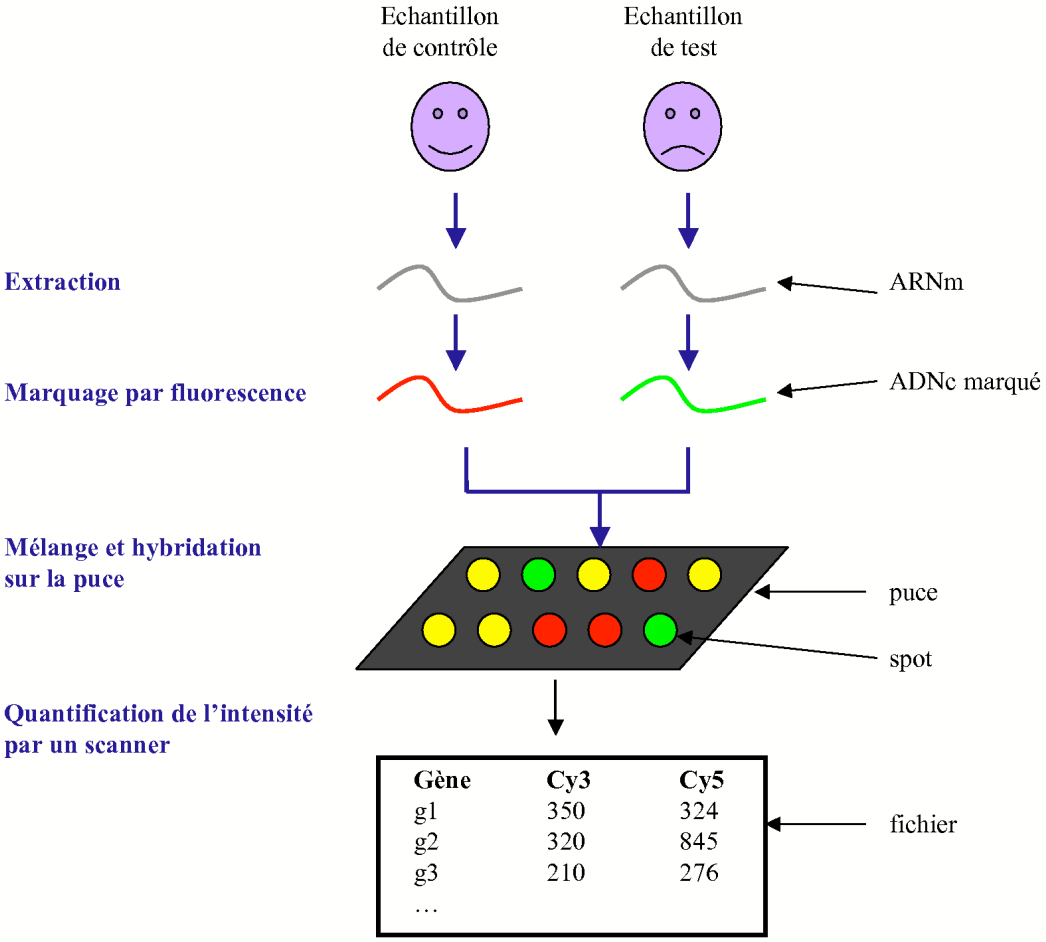


FIG. 1.1 – **Illustration des puces à ADN.** Une puce à ADN permet de refléter le niveau d'expression de milliers de gènes à un moment donné. Une librairie de clones d'ADNc distincts (ADN complémentaire) est tout d'abord amplifiée par la technique de PCR, puis fixée sur la puce. Deux échantillons d'ARNm sont ensuite extraits à partir de différentes conditions. Ces échantillons sont transformés en ADNc par la méthode de transcription inverse, puis ils sont marqués à l'aide de deux fluorochromes (Cy_3 et Cy_5). Les deux échantillons sont alors mélangés et hybridés sur la puce. La puce est ensuite scannée par un laser en utilisant différentes longueurs d'onde de manière à obtenir les intensités numériques de chaque spot.

une certaine dose et à un temps donné. Ces échantillons sont transformés en ADNc par la méthode de transcription inverse (*reverse transcription*), puis ils sont marqués à l'aide de deux fluorochromes (Cy_3 et Cy_5). Les deux échantillons sont alors mélangés et hybridés sur la puce. Ceci entraîne une compétition entre ces deux échantillons lors de l'hybridation de chaque ADNc au fragment qui lui est complémentaire. La puce est ensuite scannée par un laser en utilisant différentes longueurs d'onde, de manière à obtenir les intensités numériques de chaque spot pour chacune des deux conditions testées. De cette façon, une mesure relative à l'intensité globale d'hybridation est obtenue pour chaque élément sur la puce.

L'hypothèse sous-jacente à l'analyse des données transcriptome est que l'intensité mesurée pour chaque gène représente son niveau d'expression relatif. Ces techniques expérimentales ne permettent pas de quantifier de manière absolue le nombre d'ARNm par cellule. En revanche, elles permettent de comparer de manière quantitative le niveau d'expression des gènes dans deux conditions différentes. Pour un gène donné, ces niveaux d'expression sont regroupés pour l'ensemble des échantillons étudiés, formant ce qu'on appelle un profil d'expression. Les gènes sont ensuite comparés selon leur profil pour identifier des classes de comportements. Pour cela, il faut au préalable normaliser les niveaux d'expression pour pouvoir les comparer de manière pertinente. Il faut entre autres éliminer les mesures artéfactuelles, et ajuster l'ensemble des intensités mesurées pour pallier l'effet du marquage. En effet, les fluorochromes peuvent influencer l'efficacité du marquage et ne pas être détectés avec la même efficacité par le scanner. Après normalisation, les gènes qui sont différentiellement exprimés entre les deux échantillons sont sélectionnés.

1.1 Sélection des gènes différentiellement exprimés

L'analyse de l'expression des gènes permet entre autres d'identifier de façon rapide et systématique des marqueurs de certaines maladies. C'est pourquoi elle est utilisée particulièrement pour les études cliniques des maladies génétiques. Plus généralement, l'analyse des données transcriptome a pour objectif notamment d'identifier les gènes qui sont différentiellement exprimés entre plusieurs conditions ou d'identifier des gènes, dits marqueurs, spécifiques d'un tissu, d'un organe ou encore spécifiques d'un dysfonctionnement (maladie).

Après l'analyse réalisée par le scanner, un fichier est obtenu. Celui-ci contient une quantification du signal et du bruit de fond pour chacun des spots ; c'est ce que nous appelons les données brutes. L'identification des gènes différentiellement exprimés nécessite d'abord de normaliser ces données, de manière à faire des comparaisons pertinentes entre les différents niveaux d'expression mesurés [Speed, 2003]. Un grand nombre de méthodes de normalisation ont été développées et utilisées. Quackenbush en a fait une revue et a conclu qu'il n'y a pas une méthode universelle mais plutôt des méthodes adaptées à chaque cas [Quackenbush, 2001]. Ainsi, les différentes approches permettent d'explorer différents aspects des données. Malgré tout, le processus de normalisation est en général constitué de deux étapes principales [Quackenbush, 2002].

Dans un premier temps, il s'agit de traiter le bruit de fond. Différentes approches peuvent être adoptées, comme par exemple la soustraction du bruit de fond local [Ritchie *et al.*, 2007].

Dans un second temps, il s'agit de corriger les éventuels biais d'intensité à l'intérieur de chaque puce, par exemple en utilisant la méthode *lowess* [Cleveland et Devlin, 1988]. Cette régression linéaire, appelée *LOcally WEighted Scatterplot Smoothing*, corrige l'éventuelle dépendance systématique existant entre les rapports d'intensités lumineuses et l'intensité totale pour chaque spot. En effet, il a été montré que les mesures des rapports d'intensités pouvaient présenter une dépendance systématique en fonction de l'intensité totale [Quackenbush, 2001]. Ceci apparaît en général comme une déviation du niveau zéro pour les spots de faible ou de forte intensité (voir Figure 1.2). La normalisation *lowess* a pour objectif d'éliminer de tels biais systématiques. Pour cela, une régression linéaire locale pondérée est effectuée en fonction de $log_{10}(R * G)$ (R et G indiquent ici les valeurs de chaque canal Cy_3 et Cy_5). Cette fonction est ensuite soustraite localement à chaque point de mesure. Si on pose :

$$\begin{cases} T_i = \frac{R_i}{G_i} \\ x_i = log_{10}(R_i * G_i) \\ y_i = log_2(\frac{R_i}{G_i}) \\ T'_i = \frac{R'_i}{G'_i} \end{cases} \tag{1.1}$$

la dépendance $y(x_i)$ de $log_2(ratio)$ par rapport à $log_{10}(produit)$ est calculée par l'algorithme *locfit*, puis utilisée point par point pour corriger les valeurs mesurées, de telle sorte que :

$$log_2(T'_i) = log_2(T_i) - y(x_i) = log_2(T_i) - log_2(2^{y(x_i)}) \tag{1.2}$$

ou de façon équivalente :

$$log_2(T'_i) = log_2(T_i * \frac{1}{2^{y(x_i)}}) = log_2(\frac{R_i}{G_i} * \frac{1}{2^{y(x_i)}}) \tag{1.3}$$

Ainsi, il s'agit d'une transformation des intensités telle que :

$$\begin{cases} G'_i = G_i * 2^{y(x_i)} \\ R'_i = R_i \end{cases} \tag{1.4}$$

La régression *locfit* est basée sur le modèle linéaire suivant :

$$Y_i = \mu(x_i) + \epsilon_i \tag{1.5}$$

La fonction $\mu(x)$, supposée lisse, est estimée par ajustement à un modèle polynomial, en général linéaire ou quadratique, le long d'une fenêtre glissante. Ainsi, pour chaque point x, un critère des moindres carrés pondérés est considéré :

$$\sum_{i=1}^{n} W(\frac{x_i - x}{h}) \, (Y_i - (a_0 + a_1(x_i - x)))^2 \tag{1.6}$$

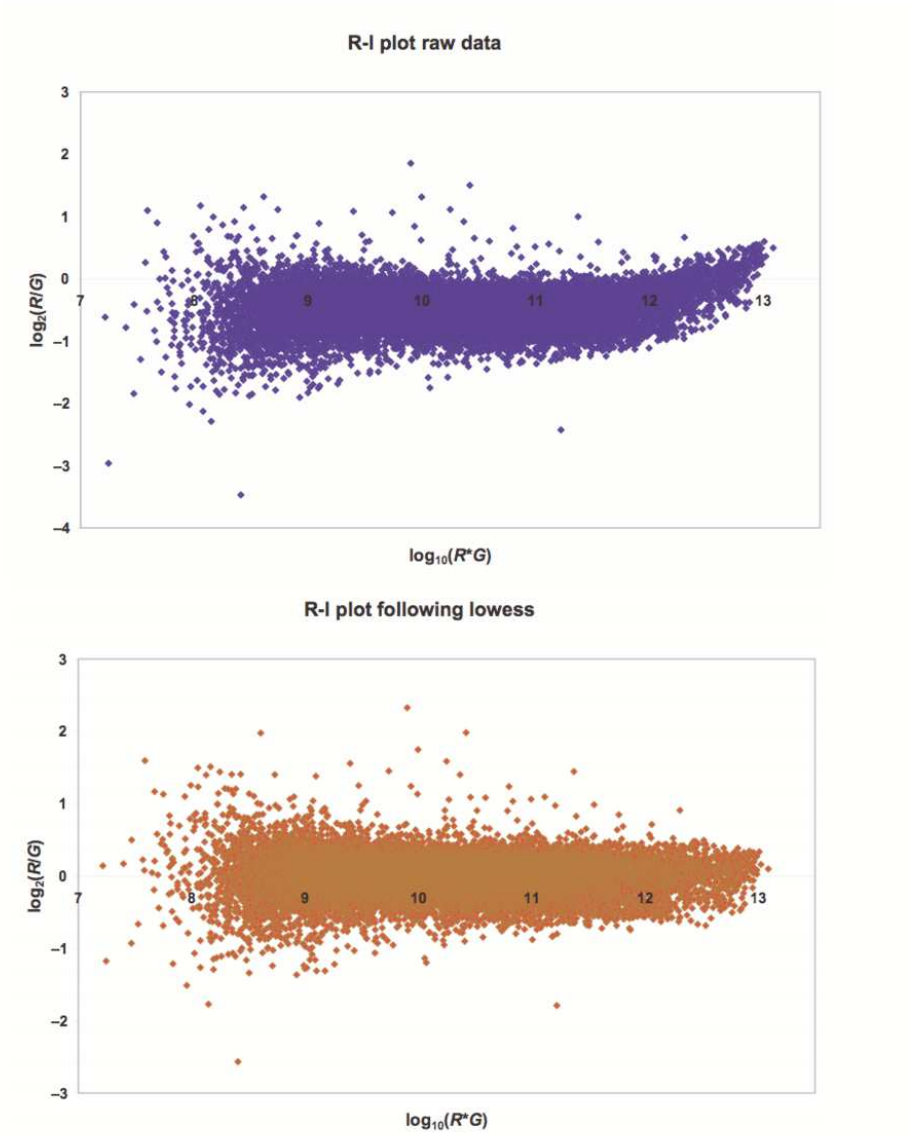


FIG. 1.2 – **Méthode de normalisation lowess.** Cette figure, extraite de la revue [Quackenbush, 2001], met en évidence l'effet de la normalisation *lowess* sur la dépendance systématique des log-ratios en fonction de l'intensité des spots. Chacun des deux graphiques représente un *R-I plot* (ratio-intensité) et permet de visualiser les effets de dépendance par rapport à l'intensité totale. Avant normalisation, nous observons une déviation par rapport au niveau zéro des log-ratios des spots de forte intensité (en-haut). Cette déviation est corrigée par la normalisation (en-bas). Les valeurs pour chaque canal sont indiquées ici par R et G.

Par défaut, la fonction de pondération est la suivante :

$$W(v) = \begin{cases} (1 - v^3)^3 & \text{si } |v| < 1 \\ 0 & \text{sinon} \end{cases} \tag{1.7}$$

La bande h contrôle le lissage de l'ajustement. Ce paramètre doit être précisé lorsqu'une normalisation *lowess* est effectuée.

Cette normalisation *lowess* est basée sur l'hypothèse que la plupart des gènes ne sont pas régulés [Speed, 2003]. Cette hypothèse est généralement vérifiée pour les études à l'échelle du génome.

Une fois que les données sont normalisées, des tests statistiques sont utilisés pour sélectionner les gènes dont le taux d'expression varie significativement entre les conditions. Ces tests sont dits paramétriques lorsqu'une hypothèse est émise sur le modèle suivi par les données, ou non paramétriques dans le cas contraire.

Dans le cas d'un test paramétrique, les taux d'expression normalisés sont supposés suivre un modèle probabiliste, comme par exemple un modèle gaussien [Long *et al.*, 2001] ou un modèle de Gamma-Gamma-Bernoulli [Newton *et al.*, 2001]. Un test statistique est ensuite appliqué, en général un *t-test* [Greller et Tobin, 1999], pour savoir dans quelle mesure les taux d'expressions d'un gène donné sont conformes au modèle. Par ailleurs, Smyth *et al.* sont partis d'un modèle paramétrique hiérarchique et l'ont développé en une approche pratique pour les expériences de transcriptome avec un nombre arbitraire de conditions et d'échantillons [Smyth, 2004]. Cette méthode a été implémentée dans le logiciel R/Bioconductor[1] [Gentleman *et al.*, 2004], [Reimers et Carey, 2006] sous la forme du package *Limma*. Son principal avantage est de pouvoir prendre en compte un faible nombre de réplicats.

De plus, il faut prendre en compte le fait qu'un grand nombre de tests statistiques sont effectués, ce qui augmente considérablement la probabilité de faire une erreur d'identification. Des travaux ont été réalisés afin de contrôler cette probabilité d'erreur. Benjamini et Hochberg ont notamment proposé une mesure appelée FDR pour *False Discovery Rate*, qui permet de contrôler le taux d'erreurs attendu [Hochberg et Benjamini, 1990]. Cette mesure est basée sur l'idée que l'on peut tolérer plus d'erreurs lorsque le nombre de tests est grand. Par exemple, un taux de cinq erreurs parmi une sélection de dix gènes est probablement trop élevé alors qu'un taux de cinq erreurs parmi une sélection de cent gènes est acceptable. Cette approche a ensuite été étendue par Reiner *et al.* notamment. Ces derniers ont utilisé le ré-échantillonnage pour améliorer les performances de la procédure initiale [Reiner *et al.*, 2003]. Une autre approche basée sur des tests de permutation a été développée par Yang *et al.* pour contrôler le FDR [Yang et Yang, 2006]. Après une comparaison de leur méthode avec notamment celle de Benjamini et Hochberg, les auteurs ont conclu qu'elle pouvait être une alternative pour trouver les gènes différentiellement exprimés. Néanmoins, ils ont souligné qu'aucune méthode ne semblait dominer les autres dans tous les cas.

Dans le cas d'un test non paramétrique, aucune distribution sous-jacente aux données

[1]The Bioconductor Project http ://www.bioconductor.org

n'est supposée au départ. Ceci peut être un avantage dans la mesure où les données transcriptome sont bruitées et ne suivent pas une loi normale en général [Hunter *et al.*, 2001].

Pour sélectionner les gènes différentiellement exprimés en utilisant des tests non paramétriques, une première approche consiste à utiliser une heuristique, c'est-à-dire une méthode approximative, en particulier en fixant des seuils sur les taux d'expression ou sur les coefficients de corrélation entre les profils d'expression. Iyer *et al.* ont par exemple considéré les gènes dont le taux d'expression était modifié d'au moins un facteur 2,2 dans au moins deux expériences [Iyer *et al.*, 1999]. DeRisi *et al.* ont également considéré un seuil, mais ils ont choisi d'identifier les gènes induits plus de deux fois par rapport au niveau moyen [DeRisi *et al.*, 1997]. Troyanskaya *et al.* ont, quant à eux, utilisé les coefficients de corrélation [Troyanskaya *et al.*, 2002]. Les auteurs ont défini un gène théorique qui leur sert de discriminant idéal entre deux classes. Ce gène théorique a un taux d'expression maximal dans un groupe d'échantillons et un taux d'expression minimal dans l'autre groupe d'échantillons. De cette façon, les gènes sélectionnés sont ceux dont le profil d'expression a le plus fort coefficient de corrélation de Pearson avec le discriminant idéal.

Voici quelques autres exemples choisis parmi les nombreuses autres méthodes possibles. Gentleman *et al.* ont par exemple utilisé un *t-test* non paramétrique associé à une correction pour les tests multiples [Gentleman *et al.*, 2005]. Par ailleurs, Tusher *et al.* ont développé la méthode SAM (Significance Analysis of Microarrays), qui assigne un score à chaque gène sur la base du changement relatif du taux d'expression par rapport à la déviation standard des différentes mesures [Tusher *et al.*, 2001]. De plus, Pan *et al.* ont développé une approche basée sur les mélanges de gaussiennes [Pan *et al.*, 2002a] et l'ont comparée à deux autres approches paramétriques, un *t-test* ordinaire et une approche par régression [Thomas *et al.*, 2001]. Ils ont conclu que ces approches donnent des résultats similaires en termes de statistiques [Pan, 2002]. D'autres tests non paramétriques ont été utilisés pour identifier les gènes différentiellement exprimés, notamment des tests basés sur les rangs (test de Wilcoxon) et des tests de permutation (test de Fisher-Pitman, test de Baumgartner-Weiss-Schindler [Neuhäuser et Senske, 2004]).

Quelle que soit la méthode utilisée pour sélectionner les gènes différentiellement exprimés, la question de la stabilité de cette sélection se pose. Un indicateur de cette stabilité est la fréquence à laquelle un gène donné est sélectionné à travers des sous-échantillons. Qiu *et al.* ont ainsi étudié la stabilité de différentes méthodes de sélection sur des données simulées et des données biologiques réelles [Qiu *et al.*, 2006]. Ils ont montré comment les méthodes de sous-échantillonnage permettent de réduire l'ensemble de gènes sélectionnés. Par ailleurs, Yang *et al.* ont proposé deux méthodes dont ils ont évalué la stabilité par du sous-échantillonnage [Yang *et al.*, 2006].

Une fois que les gènes différentiellement exprimés sont sélectionnés, il est souvent intéressant de les classer.

1.2 Classification des profils d'expression des gènes

Les expériences de transcriptome permettent de suivre le niveau d'expression des gènes au cours d'une cinétique. Ces données ont été notamment utilisées pour atteindre les trois objectifs suivants : détecter les processus cellulaires sous-jacents aux effets de régulation observés, inférer des réseaux de régulation et assigner des fonctions aux gènes. Pour cela, des méthodes de classification ont été développées. Dans le domaine de la statistique exploratoire multidimensionnelle, deux familles de méthodes se distinguent : les méthodes factorielles et les méthodes de classification [Lebart *et al.*, 1995]. Concernant les méthodes de classification, certaines utilisent une mesure de similarité entre les gènes et d'autres reposent plutôt sur des modèles statistiques [Claverie, 1999].

Une des approches consiste à identifier les gènes dont les profils sont coordonnés deux à deux. Pour cela, une distance est définie entre deux profils d'expression, comme par exemple la distance euclidienne ou le coefficient de corrélation de Pearson. Elle permet notamment de calculer une matrice de distance entre tous les profils deux à deux. Nous citons cinq exemple de méthodes de classification basées sur des distances : la classification hiérarchique [Eisen *et al.*, 1998] ; les k-moyennes [Tavazoie *et al.*, 1999] ; les cartes de Kohonen, aussi appelée *self-organising maps* (SOM) [Tamayo *et al.*, 1999], [Garrity et Lilburn, 2005] ; les machines à vecteurs de support (SVM) [Brown *et al.*, 2000] ; la classification pyramidale [Polaillon *et al.*, 2007]. A l'exception de la classification pyramidale, les classes obtenues par ces différentes méthodes sont disjointes. Pourtant, un gène peut avoir plusieurs fonctions, et par conséquent faire partie de plusieurs classes fonctionnelles. La classification pyramidale permet de tenir compte de cet aspect en proposant des classes éventuellement recouvrantes. Néanmoins, ces méthodes ont l'inconvénient majeur de ne pas tenir compte de l'ordre entre les différents points de mesure des profils d'expression et donc de négliger l'information temporelle contenue dans les données.

Une autre approche consiste à utiliser des modèles statistiques. Citons huit exemples de méthodes suivant cette approche : les splines cubiques [Bar-Joseph *et al.*, 2003] ; les courbes autorégressives [Ramoni *et al.*, 2002] ; les cinétiques du premier ordre [Sásik *et al.*, 2002] ; les modèles de markov cachés (HMM) [Ji *et al.*, 2003] ; les modèles bayésiens [Yeung *et al.*, 2005], les modèles d'inférence [Peddada *et al.*, 2003] ; les mélanges gaussiens [Medvedovic et Sivaganesan, 2002], [Pan *et al.*, 2002b] ; les contrastes linéaires [Li *et al.*, 2006]. Ces méthodes ont l'inconvénient principal de poser des hypothèses, souvent importantes, en supposant que les données suivent un modèle donné. De plus, elles ne sont que rarement adaptées au peu de données de la plupart des expériences et aux intervalles non uniformes entre les points de mesure.

Nous allons présenter plus en détail les classifications hiérarchique et pyramidale car elles ont été utilisées dans ce travail.

1.2.1 Formalisation des données

Pour caractériser les individus que l'on cherche à classifier, les formalismes suivants sont classiquement utilisés :

1. La matrice des profils ou des modalités, $X = (x_{ik})$. Cette matrice est de taille $n \times p$, où x_{ik} correspond à la valeur prise par la k^{me} variable $(k = 1, ...p)$ pour le i^{me} individu$(i = 1, ...n)$. Chaque ligne de la matrice constitue ainsi le profil, ou l'ensemble des modalités, du i^{me} individu.
2. La matrice de dissimilarité, $D = (d_{ij})$, de taille $n \times n$, à valeurs dans $\mathbb{R}_+$. Chaque valeur d_{ij} de la matrice caractérise le degré de dissimilarité entre les deux individus i et j, où $(i, j = 1, ...n)$. Par définition, la dissimilarité est positive et symétrique :
   – $(\forall i, j)\ (d_{ij} \geq 0)$
   – $(\forall i, j)\ (d_{ij} = d_{ji})$

Notons que l'on peut passer du premier formalisme au second en définissant par exemple d_{ij} comme la distance euclidienne entre deux objets i et j. De plus, toute dissimilarité étant symétrique, il est possible d'utiliser une matrice triangulaire.

1.2.2 Le modèle hiérarchique

Le modèle hiérarchique permet la représentation d'inclusions de sous-ensembles disjoints d'un groupe d'objets (voir Définition 1.1). Une hiérarchie est classiquement représentée sous la forme d'un arbre binaire (voir Figure 1.3). Les feuilles de cet arbre représentent les objets à classer (voir la Terminologie page 19 pour les termes arbre, feuille, racine). Chaque nœud h de l'arbre représente un sous-ensemble d'objets, aussi appelé classe. Ce sous-ensemble est composé par tous les objets qui sont des feuilles du sous-arbre de racine h. Par exemple, le nœud H représente le sous-ensemble $\{A, C, D\}$. La relation d'inclusion est inférente à la structure de l'arbre. Ainsi, le sous-ensemble représenté par le nœud F est inclus dans celui représenté par le nœud H. Par conséquent, la racine de l'arbre représente l'ensemble Ω composé de la totalité des objets étudiés.

Définition 1.1 (Hiérarchie) *Soit Ω un ensemble fini d'objets. (H, f) est une hiérarchie indicée sur Ω, si H est un ensemble de parties de Ω, et si f est une fonction de H à valeurs dans $\mathbb{R}_+$ satisfaisant les conditions suivantes :*

1. $\Omega \in H$
2. $\forall\ \omega \in \Omega, \{\omega\} \in H$
3. $\forall\ (h_1, h_2) \in H \times H,\ h_1 \cap h_2 \neq \emptyset \Rightarrow h_1 \subset h_2$ *ou* $h_2 \subset h_1$
4. $f(h) = 0 \Leftrightarrow |h| = 1$
5. $h_1 \subset h_2 \Leftrightarrow f(h_1) \leq f(h_2)$

Dans certains cas, cet arbre peut être valué, c'est-à-dire qu'à chaque nœud qui le compose, une valeur est associée mesurant la dispersion des objets qu'il contient. La hiérarchie est alors dite *indicée*. Pour améliorer la lisibilité de cet arbre, les nœuds sont positionnés proportionnellement à cet indice (voir Figure 1.3). Les nœuds les plus proches des feuilles sont composés d'objets proches, aussi la valeur associée à ces nœuds est faible. En revanche, pour les nœuds plus éloignés, cette valeur augmente, reflétant la plus grande dispersion des objets entre eux.

Dans une hiérarchie, il existe un ensemble d'ordres compatibles sur les objets. Un ordre est dit compatible s'il n'existe aucun croisement dans les branches de l'arbre. Ainsi, on peut passer d'un ordre compatible à un autre en effectuant des permutations des éléments appartenant à un nœud. Par exemple, sur la figure 1.4, on peut passer de l'ordre $\{A, B, C, D, E\}$ à l'ordre $\{A, B, D, E, C\}$ par des rotations des objets autour des nœuds 1 et 2.

Les algorithmes permettant de construire des hiérarchies sont nombreux [Gordon, 1996]. Celui que nous avons utilisé est la classification hiérarchique ascendante (CAH). Il est de type agglomératif : l'ensemble des nœuds composant la hiérarchie est construit progressivement. Au départ, les seuls nœuds existants sont les n objets à classer. Itérativement, ce nombre de nœuds est réduit en agglomérant les nœuds les plus proches pour en former un nouveau. Un indice est alors attribué à ce nouveau nœud en fonction de la proximité des objets qu'il contient. Ensuite, les distances entre ce nouveau nœud et les autres nœuds sont calculées. Pour cela, différents critères ont été définis de manière à calculer la distance entre un élément et une classe, mais aussi la distance entre deux classes. Ainsi, ces critères permettent de sélectionner les individus ou les classes les plus proches, comme le critère de Ward ou le critère du saut minimal, aussi appelé lien simple (*single linkage*). La méthode de Ward consiste à choisir à chaque étape le regroupement de classes tel que l'augmentation de l'indice soit minimum. Lorsque le lien simple est utilisé, des groupes sont rassemblés si un individu d'un groupe est le plus proche d'un élément du second groupe.

1.2.3 Le modèle pyramidal

Le modèle pyramidal est une généralisation du modèle hiérarchique [Bertrand et Diday, 1990]. Il permet la représentation d'inclusions de sous-ensembles d'un groupe d'objets (voir Définition 1.2). À la différence du modèle hiérarchique, ces sous-ensembles ne sont pas nécessairement disjoints.

Définition 1.2 (Pyramide) *Soit Ω un ensemble fini d'objets. (P, f) est une pyramide indicée sur Ω, si P est un ensemble non-vide de sous-ensembles d'éléments de Ω, et si f est une fonction de P dans $\mathbb{R}_+$ satisfaisant les conditions suivantes :*

1. $\Omega \in P$
2. $\forall\ \omega \in \Omega, \{\omega\} \in P$
3. $\forall\ (p_1, p_2) \in P \times P$, $p_1 \cap p_2 = \emptyset$ *ou* $p_1 \cap p_2 \in P$
4. $\exists$ *un ordre total* $\preceq$ *défini sur* Ω, *tel que tout élément de* P *est un intervalle de cet ordre*
5. $f(p) = 0 \Leftrightarrow |p| = 1$
6. $\forall\ (p_1, p_2) \in P \times P$, $p_1 \subset p_2 \Rightarrow f(p_1) \leq f(p_2)$

Une pyramide est classiquement représentée sous la forme d'un graphe planaire (voir Figure 1.5 et Terminologie page 19). À l'image des hiérarchies, les nœuds terminaux de ce graphe représentent les objets à classer. Chaque nœud interne, aussi appelé *palier*,

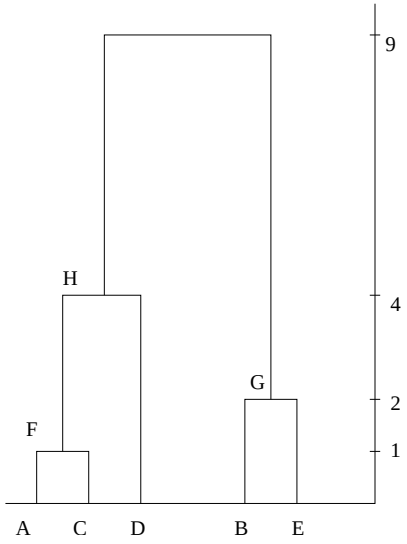


FIG. 1.3 – **Exemple d'une hiérarchie indicée.** La valeur de l'indice quantifie la distance entre les classes.

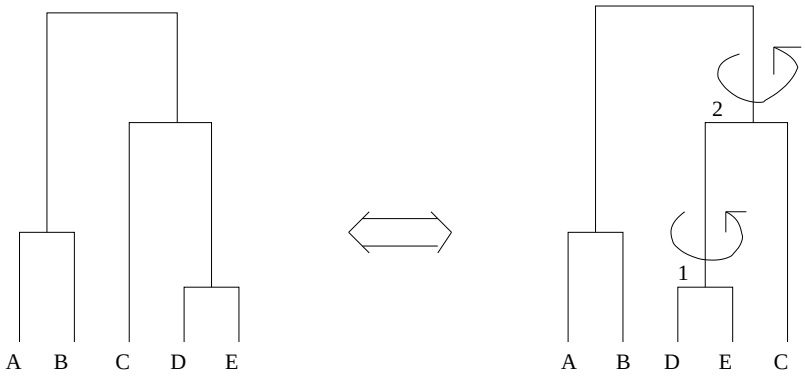


FIG. 1.4 – **Illustration de la permutation des classes d'une hiérarchie.** On peut passer ici de l'ordre $\{A, B, C, D, E\}$ à l'ordre $\{A, B, D, E, C\}$ par des rotations des objets autour des nœuds 1 et 2.

représente un sous-ensemble d'objets de Ω. À la différence des hiérarchies, les pyramides permettent de visualiser des recouvrements. En effet, un même objet peut appartenir aux deux sous-ensemble engendrés par deux paliers de même niveau. Autrement dit, en comparaison avec les arbres binaires, un nœud peut avoir deux pères. Les classes ne sont donc plus nécessairement disjointes.

Un objet peut être inclus dans au plus deux paliers ; c'est par exemple le cas de l'objet B sur la figure 1.5. Pour éviter des chevauchements qui rendraient la lecture difficile, chaque côté d'un palier est représenté de manière oblique, orienté vers l'intérieur du palier. De la même manière que pour les hiérarchies, un indice peut être associé à chaque palier d'une pyramide. La pyramide est alors dite *indicée*.

À la différence des hiérarchies, les éléments d'un palier ne peuvent pas systématiquement être permutés. En effet, un objet qui est à l'intersection de deux paliers empêche la permutation d'un de ces derniers car cette permutation engendrerait des croisements dans les branches de l'arbre, ce qui est contraire à la définition de l'ordre compatible. L'ordre engendré par la pyramide sur les objets est donc plus précis, et c'est là son avantage par rapport à la hiérarchie. On voit à titre d'illustration sur la figure 1.6 que le nœud B empêche la rotation des paliers 1 et 2.

Toutefois, ce gain en précision sur le classement des objets est compensé par une perte sur le plan de la complexité algorithmique. L'algorithme que nous avons utilisé est la classification ascendante pyramidale (CAP). Étant une adaptation de l'algorithme de CAH, il est également de type agglomératif, mais il est beaucoup plus limité sur le nombre d'individus à traiter. En effet, l'implémentation que nous avons utilisée pouvait accepter jusqu'à 250 individus pour des raisons de complexité en espace.

1.2.4 Le modèle composite de classification

Le modèle composite de classification (CCM pour *Composite Clustering Method*) [Aude, 1999] est basé sur les modèles hiérarchique et pyramidal. Ce modèle composite a été proposé dans le cadre de la classification des génomes microbiens. Son objectif est l'obtention d'un ensemble de familles de séquences similaires. L'ensemble de départ est constitué de séquences de gènes. Une méthode de comparaison de séquences est d'abord appliquée afin de transformer cet ensemble de départ en une matrice de dissimilarité. Les objets sont alors des classes de séquences.

La classification hiérarchique est utilisée pour comprendre la manière dont ces classes s'agrègent. Elle permet de mettre en évidence les relations inter-classes.

La classification pyramidale est utilisée pour analyser plus en détails chaque classe de séquences. Elle permet de visualiser la structure et les groupes à l'intérieur de chaque classe.

Nous avons exposé les modèles hiérarchique et pyramidal, ainsi que le modèle composite, dans le cadre de la classification des gènes. Nous allons maintenant aborder les relations entre le niveau d'expression et les propriétés des gènes.

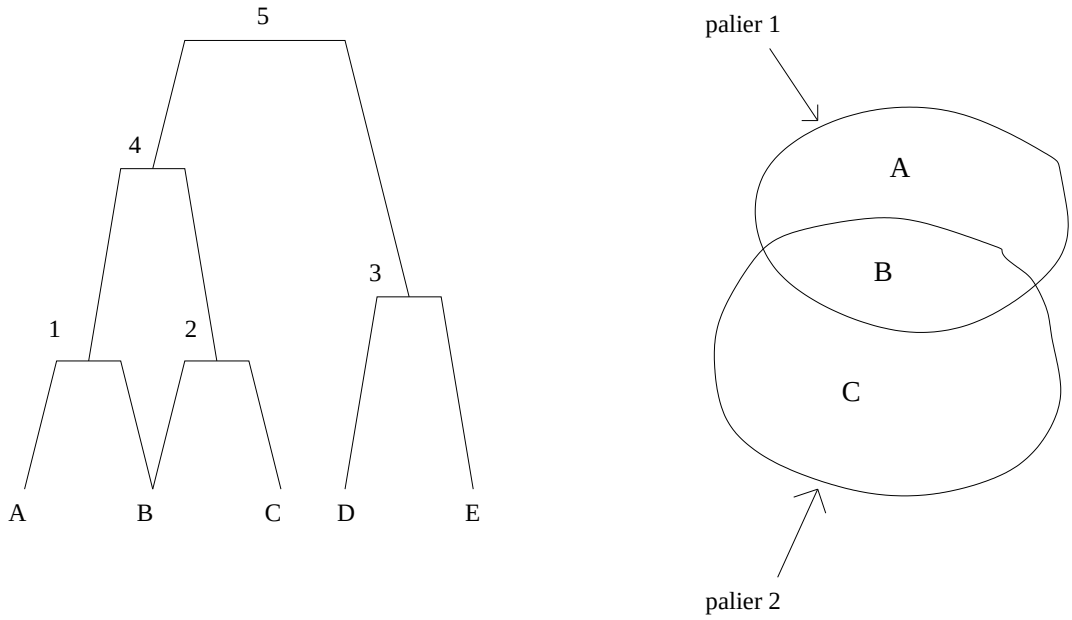


FIG. 1.5 – **Exemple d'une pyramide.** Ce schéma représente une pyramide et permet d'illustrer en particulier le phénomène de recouvrement des classes d'objets.

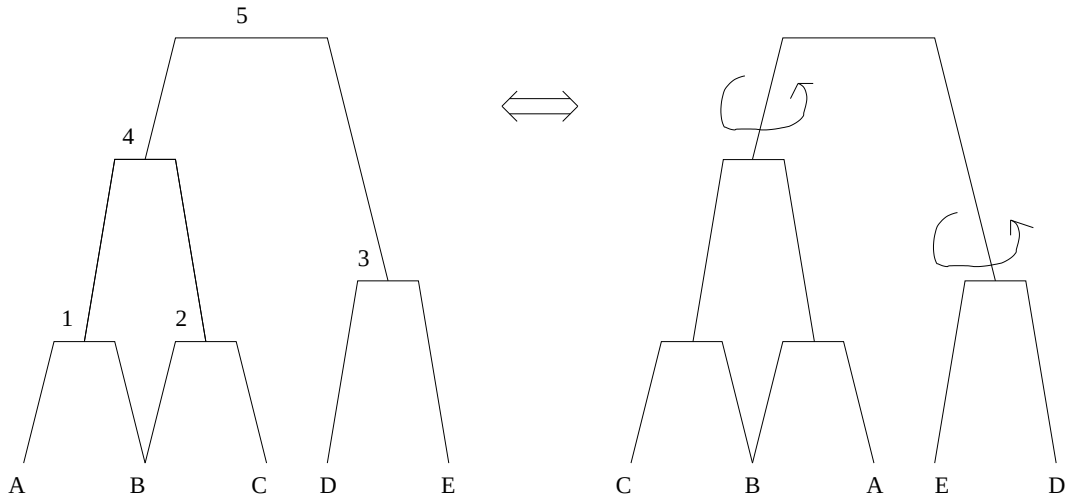


FIG. 1.6 – **Illustration de l'ordre compatible pour une pyramide.** On voit sur cette figure que le nœud B empêche la rotation des paliers 1 et 2.

1.3 Mise en évidence de relations entre les propriétés des gènes et leur niveau d'expression

Différentes études ont mis en évidence des relations entre les propriétés des gènes, ou de leurs produits comme les ARNm et les protéines, et leur niveau d'expression. En particulier, Fauchon *et al.* ont montré que le niveau d'expression de certains gènes était lié à la teneur en soufre des protéines correspondantes [Fauchon *et al.*, 2002]. Ils ont par exemple étudié *Pdc1p* et *Pdc6p*. D'un côté, *Pdc1p* est une enzyme généralement très abondante et avec une forte teneur en soufre (16 acides aminés soufrés). D'un autre côté, *Pdc6p* est très peu abondante et avec une faible teneur en soufre (5 acides aminés soufrés). Lorsque la levure est mise en présence de cadmium, le gène codant *Pdc*1*p* est largement réprimé, alors que le gène codant *Pdc*6*p* est fortement induit. Ceci peut s'expliquer par le fait que la détoxification nécessite de redirigier en grande partie le soufre vers la voie de biosynthèse du glutathion qui est utilisé pour évacuer le cadmium de la cellule.

Une autre étude a notamment porté sur le biais de codon des gènes. Rappelons d'abord ce que le biais de codon signifie. Le code génétique désigne le système de correspondance des triplets de nucléotides, appelés codons, en acides aminés. Les ribosomes traduisent ainsi, en suivant ce code, l'enchaînement des bases nucléotidiques de l'ARN en une séquence d'acides aminés au cours de la traduction. Ce code génétique est redondant, c'est-à-dire que plusieurs triplets ou codons peuvent coder pour le même acide aminé. Par conséquent, lorsque l'on considère un acide aminé donné, il peut provenir de différents codons.

La question était alors de savoir si chaque codon était également représenté dans les différents gènes ou pas. Bennetzen *et al.* ont montré que ce n'est pas le cas [Bennetzen et Hall, 1982]. Ainsi, chez *S. cerevisiae* et chez *E. coli*, la plupart des gènes codant des ARNm montrent un biais significatif dans le choix du codon parmi les différentes possibilités pour un acide aminé donné. En outre, Bennetzen *et al.* ont noté que les gènes qui sont fortement exprimés sont plus biaisés que les gènes peu exprimés.

Par la suite, Sharp *et al.* ont confirmé ces résultat chez *S. cerevisiae*. Ils ont également défini un index afin de mesurer ce biais de codon de manière quantitative [Sharp et Li, 1987]. Cet index est appelé CAI pour *Codon Adaptation Index*. Il a permis de mettre en évidence une corrélation positive entre le niveau d'expression d'un gène et son niveau de biais de codon [Jansen *et al.*, 2003a].

Finalement, Mrazek *et al.* ont identifié les gènes fortement exprimés chez *Synechocystis* en se basant sur leur biais de codon [Mrázek *et al.*, 2001].

Nous avons exposé les principales méthodes d'analyse de données transcriptome, et en particulier les étapes de normalisation et de classification. Concernant la classification, nous avons présenté en détails les modèles hiérarchique et pyramidal.
La suite de cette étude bibliographique a pour objectif d'introduire quelques notions sur l'homologie.

Chapitre 2

Présentation de la notion d'homologie

"Je crains que Homo Sapiens ne soit qu'une chose si petite dans un vaste univers, un événement évolutif hautement improbable."

Stephen Jay Gould,
La vie est belle, 1998

Nous présentons d'abord les principaux concepts liés à l'homologie dans le cadre de l'évolution des gènes et des protéines. Puis nous exposons quelques méthodes de détection des homologues.

2.1 Définitions

Dans le cadre de l'évolution des génomes, trois types d'événements majeurs peuvent avoir lieu : la duplication, le transfert et la spéciation. La duplication correspond à la copie d'une région chromosomique au sein d'un même génome. Le transfert latéral correspond à la copie d'une région chromosomique d'un organisme vers un autre, par l'intermédiaire d'un vecteur externe, comme par exemple un phage ou un plasmide. La spéciation correspond à la séparation de deux espèces à partir d'une espèce ancestrale. Différents termes ont été définis pour identifier les relations entre les gènes subissant ces événements (voir Figure 2.1).

Le terme homologie a été introduit par Owen en 1848 pour désigner la propriété d'un même organe trouvé chez différents animaux comprenant toutes les variétés de forme et de fonction [Owen, 1848]. De manière simple, deux éléments sont dits homologues s'ils partagent un ancêtre commun. Ainsi, ce terme peut s'appliquer notamment à des gènes et des protéines. L'homologie est donc la divergence à partir d'un ancêtre commun (évolution divergente). Elle se différencie en cela de l'analogie où les deux éléments analogues ont évolué séparément pour aboutir à la même fonction (évolution convergente). Dans leur revue sur l'utilisation de l'homologie en génomique comparative, Descorps-Declère *et al.* soulignent que cette notion d'homologie est une propriété de "tout-ou-rien", c'est-à-dire que deux éléments sont ou ne sont pas homologues, et rappellent qu'il s'agit toujours d'une hypothèse [Descorps-Declère *et al.*, 2008].

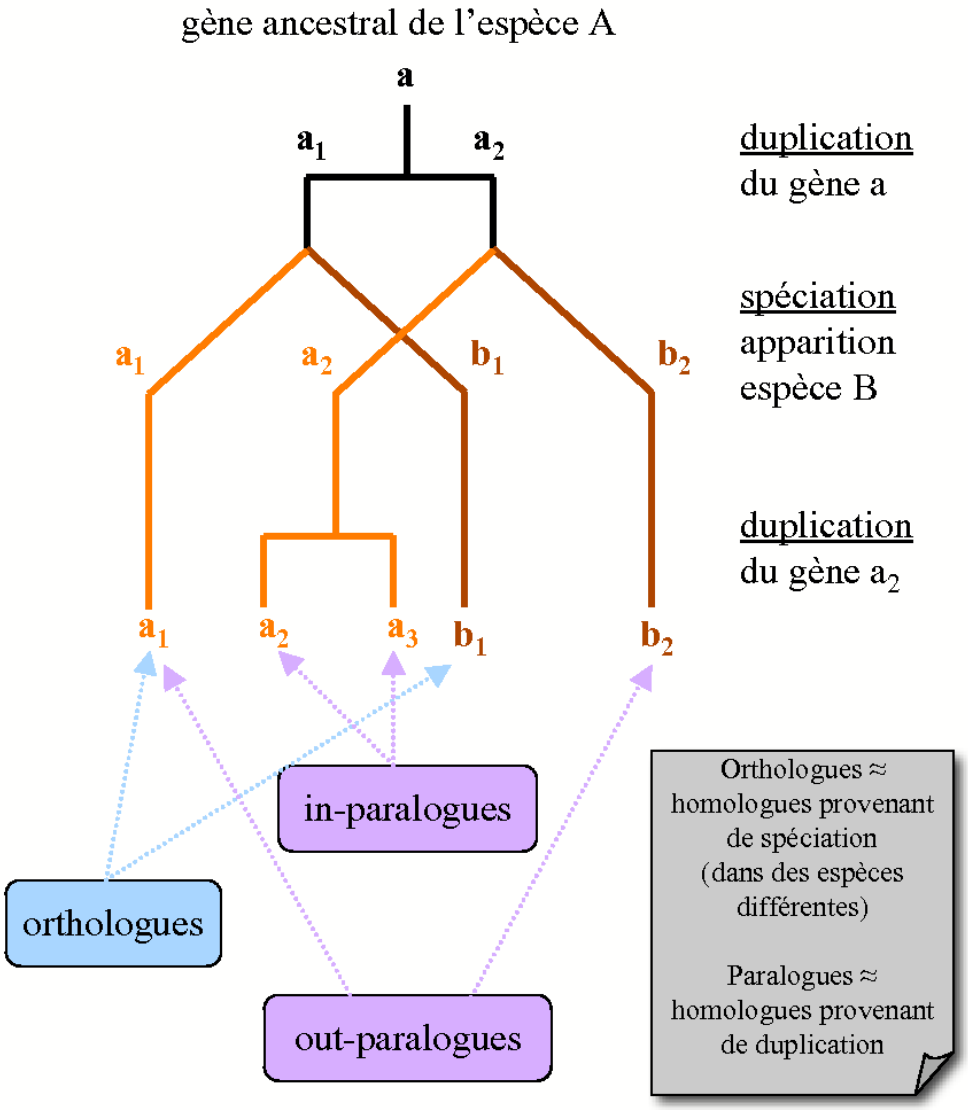


FIG. 2.1 – **Évolution des génomes.** L'évolution des génomes est constituée en particulier d'événements de spéciation et d'événements de duplication. Tous les gènes descendant d'un ancêtre commun sont dits homologues. Parmi eux, on distingue les orthologues, qui ont divergé par un événement de spéciation (a_1 et b_1) et les paralogues, qui ont divergé par un événement de duplication. Si aucun événement de spéciation n'a eu lieu depuis la duplication, on parle d'in-paralogues (a_2 et a_3). Sinon, on parle d'out-paralogues (a_1 et b_2 ou a_2 et b_1).

En 1970, Fitch *et al.* ont distingué différentes classes d'homologies [Fitch, 1970] :
- Orthologie : les gènes orthologues ont divergé par spéciation
- Paralogie : les gènes paralogues proviennent d'une duplication ancestrale
- Xénologie : les gènes xénologues ont été transférés latéralement

Les gènes paralogues ont depuis été divisés en deux catégories [Remm *et al.*, 2001] :
- in-paralogue : paralogues ayant été dupliqués récemment, après le dernier événement de spéciation
- out-paralogue : paralogues ayant été dupliqués avant un événement de spéciation

Il a été montré que les protéines orthologues ont tendance à garder la même fonction [Mushegian *et al.*, 1998]. Par contre, les protéines paralogues ont plus souvent des fonctions divergentes.

2.2 Détection des homologues

La détection de protéines homologues se fait en général en deux étapes. Tout d'abord, les protéines homologues potentielles sont détectées par une similarité de séquence forte, en générale supérieure à un seuil fixé [Doolittle, 1981]. Dans un second temps, il est nécessaire de construire un arbre reflétant l'évolution, en se basant sur des alignements multiples de séquences.

Néanmoins, les relations d'homologie ne sont pas seulement binaires. De nombreux événements de duplication et de spéciation rendent ces relations complexes. Ainsi, des groupes d'orthologues sont souvent considérés comme dans les classifications COG (Cluster of Orthologous Group)[Tatusov *et al.*, 1997], [Tatusov *et al.*, 2003] et PORC (Putative ORthologous Cluster) [Kersey *et al.*, 2005].

Nous avons présenté la notion d'homologie et les concepts qui lui sont associés. De plus, nous avons exposé quelques méthodes permettant de détecter l'homologie.
La suite de cette étude bibliographique est consacrée aux interactions protéine-protéine.

Chapitre 3

Présentation des interactions protéine-protéine

"Les organismes ne sont pas simplement élaborés selon un ensemble d'instructions. Il n'y a pas de façon simple de séparer les instructions du processus qui les exécute, ni de distinguer le plan de son exécution."

The art of gene,
Enrico Coen, 1999

L'objectif de cette section est de présenter les interactions protéine-protéine en général et notamment les méthodes expérimentales permettant de les identifier à grande échelle, l'état actuel des connaissances au niveaux des protéomes complets et enfin les méthodes de prédiction *in-silico* d'interactions protéine-protéine.

Le terme interactome a été introduit en 1999 par Sanchez *et al.* pour dénoter l'ensemble des interactions encodées par un génome, qu'il s'agisse d'interactions ADN-protéine, ARN-protéine ou protéine-protéine [Sanchez *et al.*, 1999]. Dans notre cas, nous nous limitons aux interactions protéine-protéine. Nous utilisons donc le terme interactome dans le sens actuellement largement utilisé et limité aux interactions protéine-protéine.

3.1 Identification des interactions protéine-protéine

Les interactions protéine-protéine jouent un rôle primordial dans la plupart des mécanismes cellulaires. L'identification des interactions est donc essentielle pour la compréhension des mécanismes biologiques au niveau cellulaire. Toutes ces interactions ne se produisent pas au même endroit, au même moment et avec la même force. La force d'une interaction est notamment quantifiée par une constante d'affinité. Ces constantes varient sur une échelle de six ordres de grandeur (du micromolaire au picomolaire) [Levy et Pereira-Leal, 2008]. Malgré la diversité des interactions, deux groupes sont en général considérés : les interactions permanentes et les interactions transitoires (voir la

partie haute de la Figure 3.1) [Nooren et Thornton, 2003], [Levy et Pereira-Leal, 2008]. Une interaction permanente est en général très stable et existe donc sous sa forme complexée seulement. Une interaction transitoire, quant à elle, peut s'associer et se dissocier *in vivo*. Nooren *et al.* ont proposé de distinguer deux catégories parmi les interactions transitoires : une interaction transitoire faible caractérise un équilibre dynamique en solution où l'interaction est cassée et reconstruite de manière continue ; une interaction transitoire forte est déclenchée par un processus actif comme par exemple la co-localisation ou un changement de conformation [Nooren et Thornton, 2003] (voir la Figure 3.1). Même s'il n'existe pas de frontière bien définie entre ces classes d'interactions, leurs propriétés biologiques ont été caractérisées suivant différentes tendances (voir la partie basse de la Figure 3.1).

Notons qu'une interaction protéine-protéine qui s'effectue entre deux protéines identiques est qualifiée d'homodimère. Dans le cas contraire, elle est qualifiée d'hétérodimère.

Par ailleurs, une protéine se situe en général dans un environnement rempli d'autres protéines qui sont potentiellement des partenaires d'interaction. La plupart des protéines sont très spécifiques dans leurs choix de partenaires, bien que certaines soient multi-spécifiques [Nooren et Thornton, 2003]. La spécificité provient des propriétés structurales et physico-chimiques des deux protéines, notamment de la complémentarité de forme. Cela dit, la localisation joue aussi un rôle important. Par conséquent, il est admis qu'il y a un certain degré de conservation des interactions entre des protéines similaires. En effet, il a été montré que des protéines fortement homologues interagissent en général de la même manière (voir page 43). En fait, les interactions protéine-protéine conduisent à des contraintes évolutives sur la séquence des protéines et la divergence structurale [Teichmann, 2002]. Des travaux récents ont estimé que le nombre total de types d'interactions est limité et plutôt faible, à savoir entre 6 000 et 10 000 environ [Aloy et Russell, 2004].

De nombreuses techniques expérimentales permettent de mettre en évidence des interactions protéine-protéine [Shoemaker et Panchenko, 2007a]. Nous présentons ici les deux principales approches expérimentales permettant d'identifier des interactions protéine-protéine à l'échelle du protéome, à savoir le double-hybride chez la levure (*Y2H*) et l'analyse par spectrométrie de masse de complexes purifiés (*AP-MS*).

3.1.1 Le double-hybride : *Y2H*

Le développement de la technique de *Y2H* a permis d'accélérer considérablement le criblage des interactions protéine-protéine dans le vivant [Fields et Song, 1989]. Cette technique est basée sur le fait que la plupart des activateurs de la transcription chez les eucaryotes ont au moins deux domaines distincts. Le premier domaine se lie directement à une séquence promotrice de l'ADN, il est qualifié de domaine de liaison (BD pour *Binding Domain*). Le second active la transcription, il est qualifié de domaine d'activation (AD pour *Activation Domain*). Il a été montré que la séparation des domaines BD et AD inactive la transcription. Néanmoins, la transcription peut être réactivée si le domaine BD est proche du domaine AD.

Dans le cas du test d'une interaction entre deux protéines A et B, la méthode *Y2H*

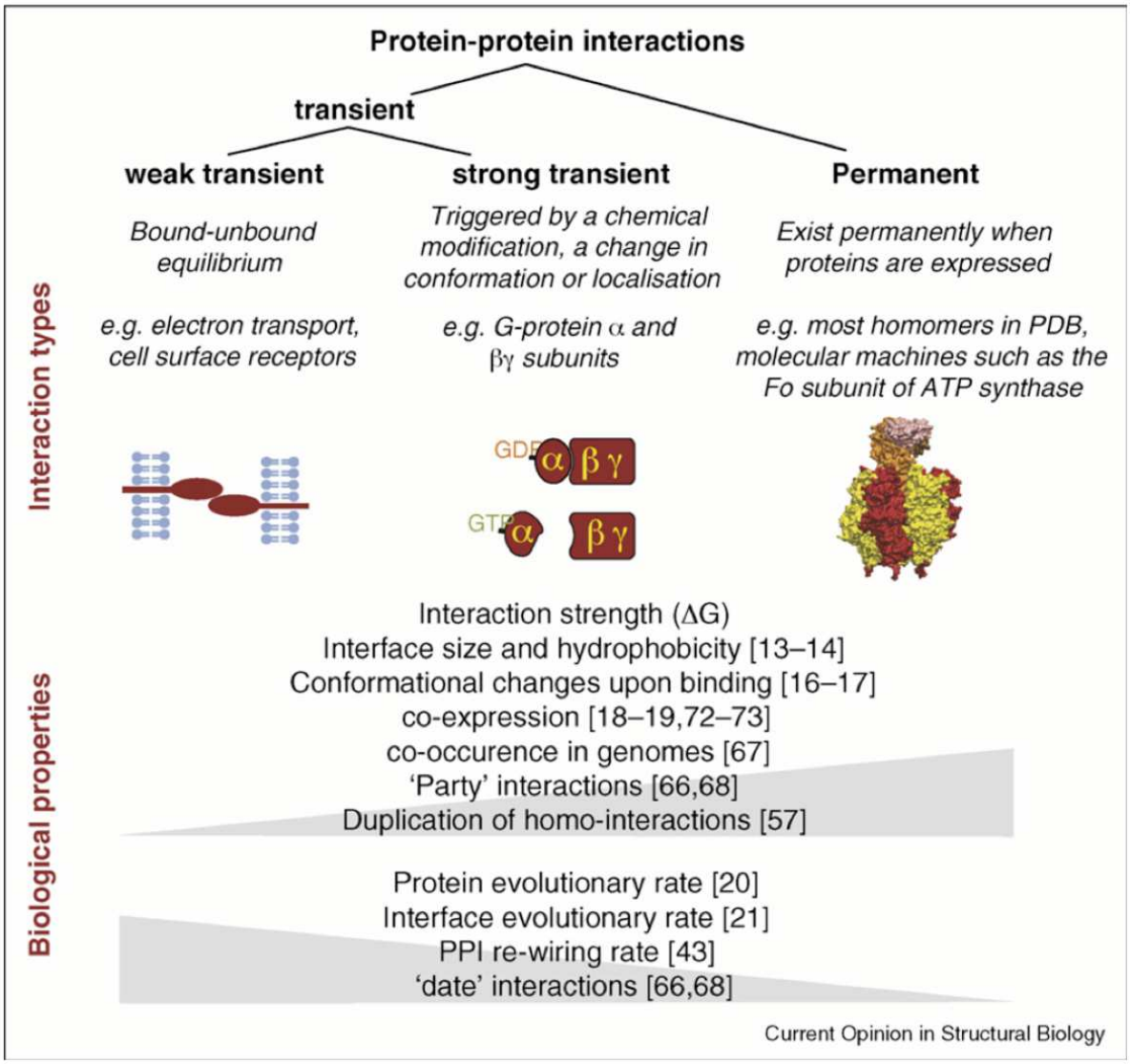


FIG. 3.1 – **Classification des interactions protéine-protéines.** La figure représente une décomposition schématique des interactions protéine-protéine en trois classes et leurs propriétés biologiques associées. Cette image est extraite de l'article [Levy et Pereira-Leal, 2008].

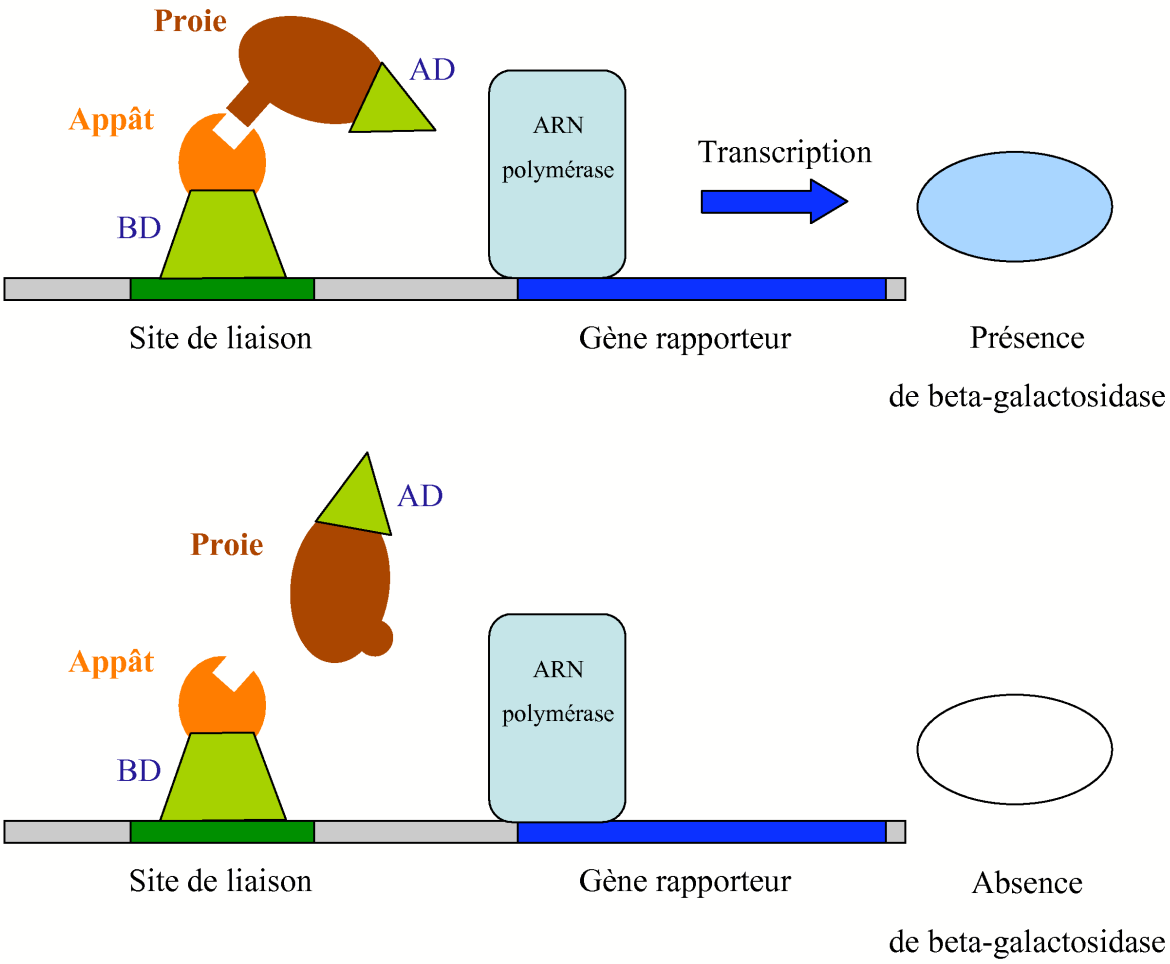


FIG. 3.2 – **Techniques d'identification des interactions protéine-protéine par *Y2H*.** Pour tester l'existence d'une interaction entre deux protéines A et B, la méthode *Y2H* est la suivante : la protéine A est fusionnée avec le domaine de liaison BD, il s'agit de la protéine appât ; la protéine B est fusionnée avec le domaine d'activation AD, il s'agit de la protéine proie. Le gène de la protéine appât est cloné dans un plasmide d'expression qui est ensuite transféré dans une cellule de levure. Si les protéines appât et proie interagissent, les domaines BD et AD vont pouvoir être proches physiquement, et ainsi activer la transcription du gène rapporteur. Cette figure est une adaptation de [Klipp *et al.*, 2005].

est la suivante : la protéine A est fusionnée avec le domaine BD, il s'agit de la protéine appât (*bait*) ; la protéine B est fusionnée avec le domaine AD, il s'agit de la protéine proie (*prey*). Après fusion, les protéines sont qualifiées de chimériques. Le gène de la protéine chimère appât est cloné dans un plasmide d'expression qui est ensuite transféré dans une cellule de levure.

Si les protéines A et B interagissent, les domaines BD et AD vont pouvoir être proches physiquement, et ainsi activer la transcription d'un gène appelé gène rapporteur (voir Figure 3.2). L'un des systèmes les plus couramment utilisés est le système GAL4, où la protéine GAL4 contrôle l'expression, chez la levure, du gène LacZ codant la beta-galactosidase. Ainsi, la présence d'une interaction entre les protéines A et B a pour conséquence la transcription du gène rapporteur et donc la production de beta-galactosidase. Suite à une réaction enzymatique, cette protéine peut ensuite être détectée par la coloration qu'elle donne aux colonies de cellules.

Pour le criblage de protéomes entiers, deux approches ont principalement été développées : une approche matricielle et une approche par criblage de banque [Legrain *et al.*, 2001].

En ce qui concerne l'approche matricielle, une liste de protéines proies et une liste de protéines appâts sont fixées au départ. Pour l'ensemble des protéines proies, un ensemble de clones est créé. Chacun des ces clones exprime une proie particulière à un endroit donné d'une plaque. Une plaque est ensuite attribuée à chacune des protéines appâts, permettant de visualiser pour cette protéine appât l'ensemble des proies qui interagissent avec elle.

En ce qui concerne l'approche par criblage, une banque ou une librairie de clones est donnée au départ. Cette banque contient des fragments aléatoires d'ADNc (ADN complémentaire) ou de cadres ouverts de lectures (ORF pour *Open Reading Frame*). Les paires positives sont sélectionnées par leur capacité à se développer sur des substrats spécifiques. Les proies sont déterminées par le séquençage d'un certain nombre de clones.

Au sein de l'approche *Y2H*, un grand nombres de techniques ont été développées dont les protocoles expérimentaux varient légèrement ou sont adaptés à certains problèmes spécifiques. L'utilisation de librairies de fragments de domaines d'activation permet notamment d'obtenir une meilleure sensibilité dans la détection des interactions protéine-protéine [Fromont-Racine *et al.*, 2002], [Boxem *et al.*, 2008].

Certaines études ont mis en évidence dans les résultats de ces approches un grand nombre de faux-positifs, c'est-à-dire des interactions détectées mais ne se produisant pas en réalité [von Mering *et al.*, 2002], ainsi que de faux-négatifs, c'est-à-dire des interactions se produisant mais n'ayant pas été détectées [Vidal et Legrain, 1999]. Ainsi, il a été montré que plusieurs études expérimentales utilisant la même technique chez la même espèce identifiaient moins de 10% d'interactions en commun par rapport à l'ensemble des interactions détectées [Ito *et al.*, 2002], [Goll et Uetz, 2006].

Néanmoins, des travaux récents ont permis de mettre en évidence la qualité des résultats obtenus par les approches *Y2H* [Yu *et al.*, 2008]. Si le taux de faux négatifs est effectivement important, ce n'est pas le cas du taux de faux positifs. Ainsi, le faible recouvrement des différents jeux de données s'explique principalement par la faible couverture

des différentes études [Gandhi *et al.*, 2006].

3.1.2 La spectrométrie de masse de complexes purifiés : *AP-MS*

La technique *AP-MS* permet de caractériser les complexes se formant autour d'une protéine appât donnée [Mann *et al.*, 2001]. Pour étudier une protéine donnée, il est souvent nécessaire de la purifier c'est-à-dire de l'isoler. Dans le cas de l'*AP-MS*, ce sont les complexes contenant la protéine appât qui vont être purifiés (voir Figure 3.3). Pour cela, la protéine appât est marquée. Les différents complexes contenus dans la cellule sont ensuite liés à une colonne d'affinité s'ils contiennent la protéine appât marquée. Par la suite, les composantes de chaque complexe sont séparées par électrophorèse. Ensuite chaque composante est découpée en fragments. Ces fragments sont caractérisés par spectrométrie de masse. Ceci permet de déterminer les protéines faisant partie d'un complexe contenant la protéine appât. L'ensemble de ces protéines identifiées sont qualifiées de protéines proies.

La remarque précédente sur les erreurs de détection et le faible recouvrement entre différentes études reste valable pour cette technique également [Ito *et al.*, 2002]. Des travaux récents du groupe de Marc Vidal ont mis en évidence les natures différentes des réseaux d'interactions protéine-protéine obtenus par ces deux approches toutes les deux utiles à l'obtention d'un réseau complet [Yu *et al.*, 2008]. Il est nécessaire d'utiliser des approches complémentaires pour obtenir un réseau le plus complet possible [Stelzl et Wanker, 2006].

Ces deux techniques expérimentales de mise en évidence d'interactions protéine-protéine utilisent des protéines appâts et des protéines proies. Cependant, la technique *Y2H* identifie des interactions binaires alors que la technique *AP-MS* identifie des complexes. Le *Y2H* a l'avantage d'être une technique *in-vivo*. De plus, elle peut identifier des interactions même transitoires. Toutefois, elle a l'inconvénient de détecter l'interaction physique de manière indirecte par un test génétique.

Ces deux approches ont l'inconvénient de détecter les interactions protéine-protéine en-dehors de leur contexte naturel. Des méthodes de détection ont été développées récemment pour identifier les interactions protéine-protéine *in vivo* et à grande échelle [Tarassov *et al.*, 2008]. D'autres promettent une meilleure détection des complexes [Heck, 2008].

3.2 Modélisation des interactions protéine-protéine

Les listes d'interactions protéine-protéine sont en général considérées sous la forme d'un réseau d'interactions protéine-protéine. Ceci permet d'avoir une vision plus globale de l'ensemble des interactions considérées et de leur organisation les unes avec les autres. Ces réseaux sont fréquemment représentés sous la forme de graphes où les nœuds représentent les protéines et les arêtes représentent les interactions entre ces protéines (voir Figure 3.4). En général, les arêtes ne sont pas orientées, ce qui correspond au concept symétrique des interactions dans la réalité. En effet, si une protéine A interagit avec une protéine B, alors la protéine B interagit avec la protéine A. Néanmoins, certaines

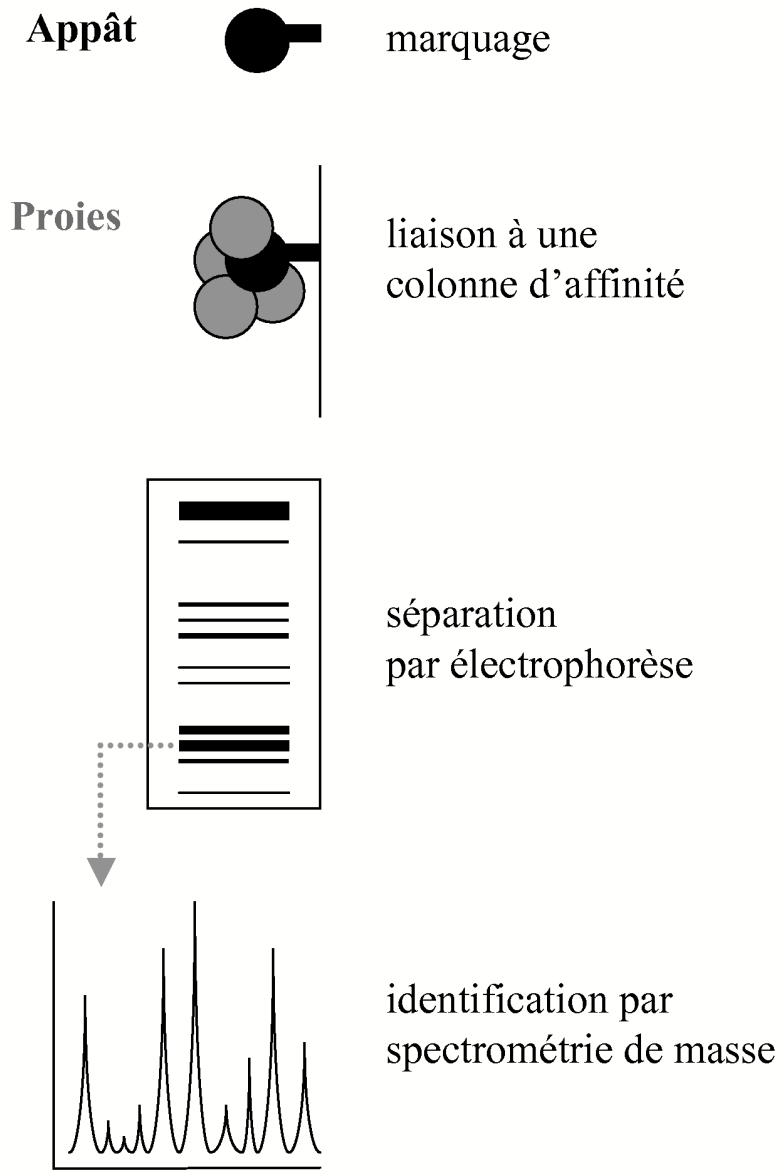


FIG. 3.3 – **Techniques d'identification des interactions protéine-protéine par *AP-MS*.** La protéine appât est d'abord marquée. Les différents complexes contenus dans la cellule sont ensuite liés à une colonne d'affinité s'ils contiennent la protéine appât marquée. Par la suite, les composantes de chaque complexe sont séparées par électrophorèse. Ensuite chaque composante est découpée en fragments. Ces fragments sont caractérisés par spectrométrie de masse. Ceci permet de déterminer les protéines faisant partie d'un complexe contenant la protéine appât. L'ensemble de ces protéines identifiées sont qualifiées de protéines proies.

méthodes d'identification expérimentales d'interactions apportent une certaine asymétrie dans la mesure où les protéines testées ont des rôles différents, appât ou proie (voir page 47). Dans ce cas, il peut être pertinent de représenter les interactions détectées sous la forme d'un graphe orienté (voir Figure 3.5).

Les interactions homodimériques, c'est-à-dire ayant lieu entre deux protéines identiques (voir page 47), sont représentées par une boucle sur le graphe.

Des réseaux d'interactions protéine-protéine, aussi appelés cartes d'interactions, ont été créés pour différents organismes modèles [Hunter, 2008], grâce à l'identification des interactions à l'échelle du protéome. De nombreuses études ont d'abord été effectuées chez *Saccharomyces cerevisiae* [Uetz *et al.*, 2000], [Ito *et al.*, 2001], [Ho *et al.*, 2002], [Gavin *et al.*, 2002], mais aussi chez *Helicobacter pylori* [Rain *et al.*, 2001]. D'autres études ont été réalisées ensuite chez *Drosophilia melanogaster* [Giot *et al.*, 2003], [Stanyon *et al.*, 2004], [Formstecher *et al.*, 2005], ainsi que chez *Caenorhabditis elegans* [Li *et al.*, 2004]. L'évolution des techniques a ensuite permis d'étudier le réseau d'interactions protéine-protéine chez *Homo sapiens* [Stelzl *et al.*, 2005], [Rual *et al.*, 2005] , dont la taille a été récemment estimée à environ 650 000 interactions [Stumpf *et al.*, 2008]. Peu d'espèces ont plus de 10 000 interactions identifiées (voir Figure 3.6).

Ces interactions protéine-protéine sont stockées dans des bases de données dont les principales sont BIND [Bader et Hogue, 2000], [Bader *et al.*, 2001], [Bader *et al.*, 2003], BioGrid [Stark *et al.*, 2006], [Breitkreutz *et al.*, 2007], DIP [Xenarios *et al.*, 2000] [Xenarios *et al.*, 2001], [Salwinski *et al.*, 2004], IntAct [Hermjakob *et al.*, 2004b], [Kerrien *et al.*, 2007a] et MINT [Zanzoni *et al.*, 2002].

3.3 Prédiction des interactions protéine-protéine

Un grand nombre de méthodes ont été développées pour prédire des interactions protéine-protéine. Des revues ont été réalisées afin de faire le point sur ces méthodes, ainsi que sur les approches expérimentales utilisées [Szilágyi *et al.*, 2005], [Shoemaker et Panchenko, 2007b].

3.3.1 Méthodes de conservation du contexte génomique

L'analyse comparative des génomes, et en particulier de la conservation des contextes génomiques à travers les espèces, a permis de mettre en évidence des liens fonctionnels entre des gènes ou entre les protéines que ces derniers codent. Ces interactions fonctionnelles ne sont pas nécessairement des interactions physiques. Différentes méthodes ont été comparées par Huynen *et al.* [Huynen *et al.*, 2000].

3.3.1.1 Transfert par interologue

La méthode de transfert par interologue est basée sur l'hypothèse que des protéines liées fonctionnellement ont tendance à co-évoluer. L'idée est donc de combiner des interactions connues dans un organisme donné et de tenter de les transférer en considérant les relations d'orthologie entre les deux organismes considérés (voir page 43). Ce concept,

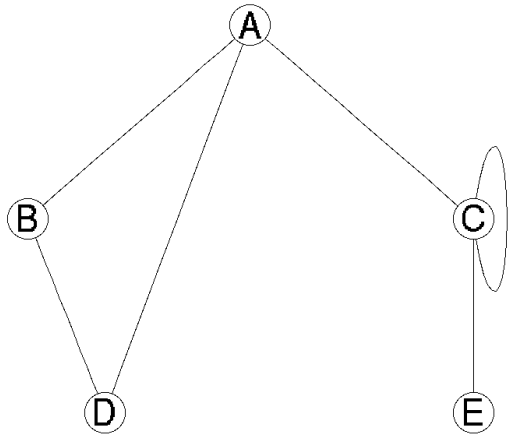


FIG. 3.4 – **Représentation des interactions protéine-protéine sous forme d'un graphe non orienté.** Chaque nœud du graphe représente une protéine. Les arêtes représentent les interactions entre les protéines.

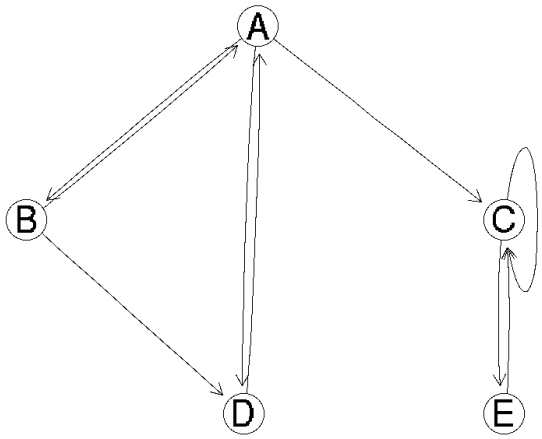


FIG. 3.5 – **Représentation des interactions protéine-protéine sous forme d'un graphe orienté.** Chaque nœud du graphe représente une protéine. Les interactions entre les protéines sont ici représentées par des arêtes orientées, ce qui peut être utile lorsque les détections des interactions A $\rightarrow$ B et B $\rightarrow$ A sont distinctes.

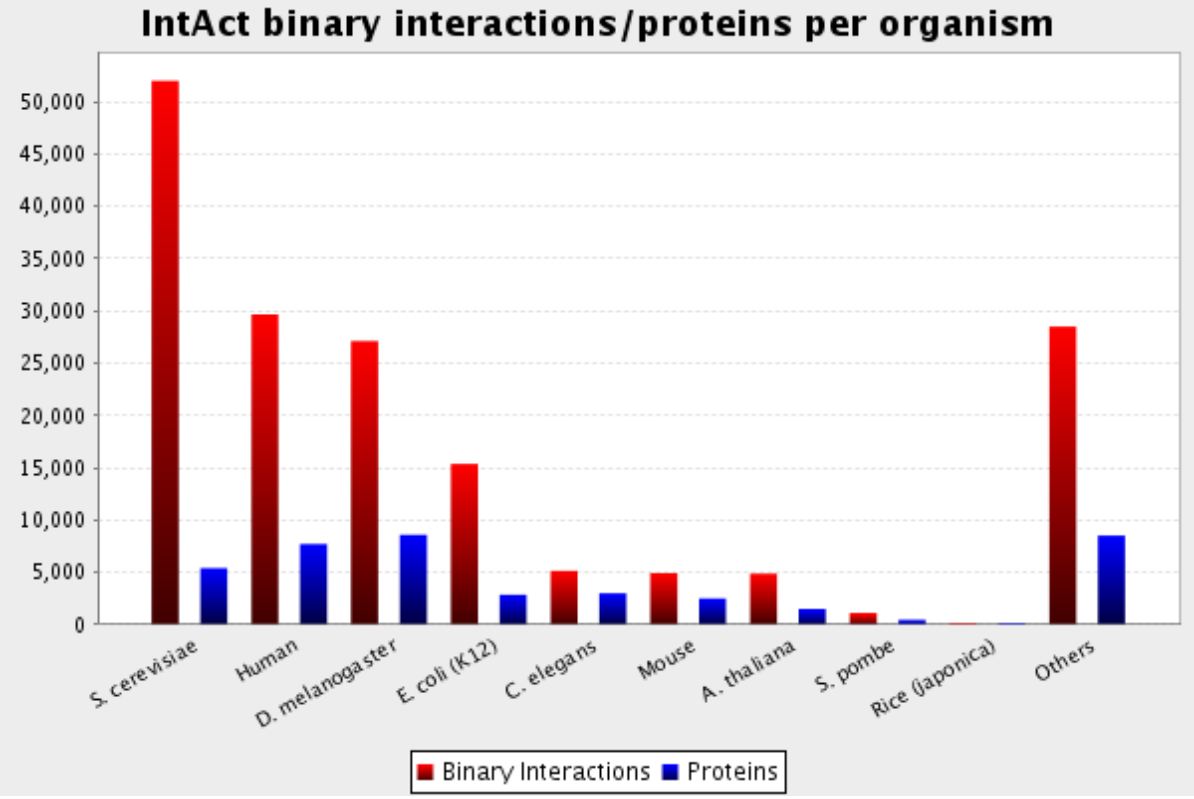


FIG. 3.6 – **Nombre d'interactions disponibles dans IntAct.** Cet histogramme représente le nombre d'interactions binaires (en rouge) présentes par espèce dans la base de données IntAct [Kerrien *et al.*, 2007a], ainsi que le nombres de protéines (en bleu) impliquées dans ces interactions. La levure est l'organisme pour lequel le plus d'interactions ont été identifiées (environ 50 000 interactions protéine-protéine). Seuls quatre organismes ont plus de 10 000 interactions disponibles : la levure (*Saccharomyces cerevisiae*), l'homme (*Homo sapiens*), la mouche du vinaigre (*Drosophilia melanogaster*) et la bactérie *Escherichia coli*.

introduit par Walhout *et al.*, est connu sous le nom d'interologue (association d'interactions et d'orthologues) [Walhout *et al.*, 2000].

De tels transferts ont déjà été effectués pour un petit nombre d'organismes modèles et avec différentes méthodes d'identification des relations d'orthologie. Matthews *et al.* ont notamment transféré deux expériences de double-hybride à grande échelle de la levure vers l'homme [Matthews *et al.*, 2001]. Des carte d'interactions protéine-protéine ont été construites pour différents organismes comme *C. elegans*, *H. pylori* ou *D. melanogaster* [Yu *et al.*, 2004b] en se basant sur l'interactome de la levure. Des réseaux d'interactions protéine-protéine ont été inférés chez l'homme à partir de plusieurs organismes sources en utilisant l'algorithme InParanoïd [Remm *et al.*, 2001] pour déterminer les protéines orthologues [Huang *et al.*, 2007b], [Huang *et al.*, 2004], [Lehner et Fraser, 2004], [Persico *et al.*, 2005]. Brown *et al.* ont développé la base de données OPHID [Brown et Jurisica, 2005] qui contient des interactions protéine-protéine chez l'homme également. Pour identifier les relations d'orthologie, les auteurs ont utilisé BLASTP [Altschul *et al.*, 1990] et l'approche dite RBH pour *Reciprocal Best Hit* [Tatusov *et al.*, 1997], [Hirsh et Fraser, 2001], [Jordan *et al.*, 2002]. Une carte d'interactions protéine-protéine a été établie chez *Plasmodium falciparum* [Wuchty et Ipsaro, 2007]. Wojcik *et al.* ont, quant à eux, construit une carte d'interactions protéine-protéine chez *Helicobacter pylori* en se basant sur les interactions domaine-domaine [Wojcik *et al.*, 2002].

3.3.1.2 Profils phylogénétiques

De la même façon que pour le transfert par interologue, cette méthode est basée sur l'hypothèse que des protéines liées fonctionnellement ont tendance à co-évoluer. Au lieu de transférer les interactions d'un organisme vers un autre en considérant des relations d'orthologie de l'un vers l'autre, on se propose ici de considérer des groupes d'organismes. Dans ce cas, des protéines reliées fonctionnellement doivent avoir des protéines homologues dans des ensembles proches d'organismes (voir page 43).

Chaque protéine est alors représentée par un vecteur binaire indiquant la présence ou l'absence de celle-ci dans chaque génome étudié [Pellegrini *et al.*, 1999], [Gertz *et al.*, 2003], [Wu *et al.*, 2003]. Les protéines ayant un profil phylogénétique proche sont prédites comme fonctionnellement liées, c'est-à-dire qu'elles participent à un même complexe protéique ou à une même voie métabolique.

Des développements ont été faits pour quantifier le niveau de confiance porté à ces prédictions [Wu *et al.*, 2006b]. Wu *et al.* ont en effet étendu la méthode initiale en prenant en compte la probabilité qu'un degré de similarité arbitraire donné entre deux profils apparaisse aléatoirement. Certaines méthodes prennent en compte également les relations phylogénétiques entre les organismes. Ceci donne des résultats plus précis mais engendre une plus forte complexité algorithmique. Cokus *et al.* ont proposé une heuristique pour prendre en compte ces relations phylogénétiques entre les organismes de manière efficace [Cokus *et al.*, 2007].

Un des principaux inconvénients de cette méthode est qu'il faut considérer des génomes entiers de façon à être sûr de l'éventuelle absence de protéines homologues pour

une protéine donnée dans ce génome. De plus, comme pour la méthode des interologues, des seuils arbitraires sont utilisés pour déterminer si une protéine homologue est présente ou non. Enfin, il a été montré que des protéines homologues peuvent avoir des fonctions différentes [Bandyopadhyay *et al.*, 2006]. Ainsi, la présence d'une protéine homologue ne garantit pas que la fonction soit conservée à travers les espèces.

3.3.1.3 Conservation du contexte génomique local

Cette méthode est basée sur le fait que, dans les génomes bactériens et archaebactériens, les gènes voisins ont tendance à coder des protéines qui montrent des interactions physiques ou fonctionnelles entre elles. Cette observation a donné naissance à plusieurs variantes.

La conservation du contexte génomique à travers les génomes peut être détectée entre autres par l'analyse de l'ordre des gènes ou de la structure en opéron [Dandekar *et al.*, 1998]. Cette conservation peut aussi être détectée par l'analyse de clusters particuliers de gènes. Overbeek *et al.* ont notamment définis des clusters comme des ensembles de gènes qui apparaissent sur le même brin d'ADN et sont séparés d'au plus 300 paires de bases [Overbeek *et al.*, 1999].

La principale limitation de ces méthodes est qu'elles ne sont pas applicables aux eucaryotes où, à part quelques exceptions, les gènes semblent être distribués aléatoirement [von Mering et Bork, 2002].

3.3.1.4 Analyse de la fusion des gènes

Une interaction fonctionnelle peut également être inférée par la présence, dans un organisme, de protéines ayant des homologues fusionnées en une seule protéine dans un autre organisme (voir page 43). L'existence d'une telle protéine de fusion dans un génome, appelée "Rosetta Stone sequence" [Marcotte *et al.*, 1999a] ou "protéine composite" [Enright *et al.*, 1999], permet de prédire une interaction entre les protéines possédant un unique domaine dans d'autres génomes même si elles ne sont pas codées par des gènes voisins. Cette méthode est limitée par le nombre d'événements de fusion de gènes qui varie selon l'organisme et les types de gènes.

Ainsi, la conservation du contexte génomique permet d'identifier des relations fonctionnelles entre des gènes, et par la suite entre des protéines, en considérant l'ordre, la proximité ou l'existence des gènes dans différentes espèces. Ceci peut s'étudier également de façon plus détaillée au niveau des bases azotées elles-mêmes dans le cas des gènes, ou au niveau des acides aminés dans le cas des protéines. C'est ce que nous allons voir maintenant.

3.3.2 Méthodes de co-évolution

Les protéines qui interagissent physiquement évoluent en général de manière coordonnée, conservant ainsi les contacts entre elles [Pazos *et al.*, 1997]. Ainsi, les méthodes

basées sur ce principe sont susceptibles de prédire des interactions pas seulement fonctionnelles mais vraiment physiques.

3.3.2.1 Similarité des arbres phylogénétiques

Les arbres phylogénétiques des protéines en interaction ont un degré de similarité plus élevé que ceux obtenus à partir de protéines qui n'interagissent pas. Goh *et al.* ont évalué la similarité des arbres phylogénétiques des deux domaines d'une protéine (la phosphoglycereate kinase) par la corrélation linéaire entre les matrices de distance utilisées pour construire les arbres [Goh *et al.*, 2000]. Pazos et Valencia ont étendu l'approche en proposant une évaluation statistique plus rigoureuse [Pazos et Valencia, 2001].

3.3.2.2 Mutations corrélées (double-hybride *in-silico*)

Pazos *et al.* ont analysé les mutations corrélées apparaissant dans des alignements multiples [Pazos *et al.*, 1997]. Ceci leur a permis de détecter des groupes de résidus qui interagissent au niveau des interfaces des protéines. Ce travail a donné naissance, par la suite, à une approche appelée "méthode de double-hybride *in silico*" [Pazos et Valencia, 2002]. L'accumulation différentielle de mutations corrélées permet d'identifier des partenaires potentiels, ainsi que les régions des séquences qui sont en interaction. Par ailleurs, Hakes *et al.* ont analysé la coévolution des protéines en interaction chez la levure, et ont mis en évidence une évolution corrélée des protéines en interaction chez les eucaryotes [Hakes *et al.*, 2007]. L'origine de cette évolution corrélée est encore largement discutée : elle peut provenir d'une co-evolution des protéines soumises aux mêmes facteurs extérieurs et/ou d'une co-adaptation, c'est-à-dire un phénomène de compensation entre les protéines en interaction [Pazos et Valencia, 2008].

Plus récemment, la méthode des mutations corrélées a été étendue par Burger *et al.* [Burger et van Nimwegen, 2008]. Au lieu de considérer des paires de protéines et leurs orthologues, Burger *et al.* effectuent des alignements multiples sur des familles entières de protéines comprenant à la fois les paralogues et les orthologues. De plus, ils utilisent un réseau bayésien pour modéliser toutes les probabilités jointes des séquences en acides aminés lors du calcul des alignements multiples.

3.3.2.3 Co-évolution de l'expression des gènes

Il a été observé que des protéines en interaction sont fréquemment co-exprimées, ceci ayant pour conséquence le maintien de la stoechiométrie entre les partenaires en interaction. À partir de cette observation, Fraser *et al.* ont montré que la co-évolution de l'expression peut être utilisée pour prédire des interactions protéine-protéine [Fraser *et al.*, 2004]. Ils ont trouvé que la co-évolution de l'expression chez la levure était un moyen plus performant de prédire des interactions physiques que la co-évolution de la séquence en acides aminés.

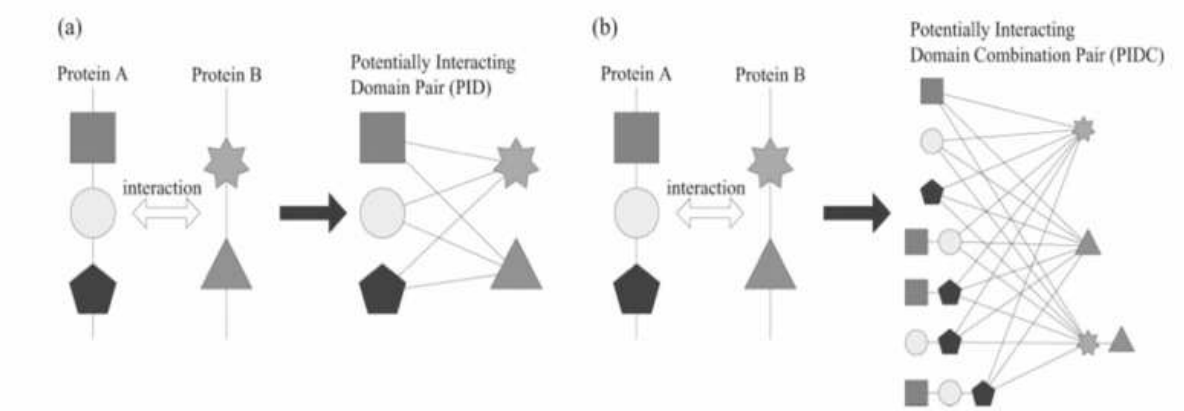


Figure 1. (**a**) A conventional domain pair based prediction model versus (**b**) proposed domain combination pair based new prediction model.

FIG. 3.7 – **Prédictions basées sur les domaines.** Le schéma illustre comment, à partir d'une interaction protéine-protéine, les potentielles interactions domaine-domaine sont construites. La première approche (à gauche) considère les différentes paires de domaines entre les deux protéines. La seconde approche (à droite) considère en plus des interactions entre des combinaisons de domaines. Cette figure provient de l'article [Han *et al.*, 2004a].

3.3.3 Méthodes basées sur les domaines

D'autres méthodes de prédiction ont été développées en se basant sur les domaines des protéines. Les domaines sont considérés comme les entités élémentaires de construction des protéines. Ce sont des unités structurales et/ou fonctionnelles qui sont conservées au cours de l'évolution. Chaque domaine contribue à la structure globale de la protéine et à ses différentes fonctions. L'hypothèse que les protéines interagissent entre elles par l'intermédiaire de leurs domaines est largement acceptée. L'idée est alors d'inférer des informations sur les interactions domaine-domaine en se basant sur les interactions protéine-protéine, puis de prédire des interactions protéine-protéine à partir des ces interactions domaine-domaine inférées.

L'un des premiers travaux dans ce domaine est la méthode d'association proposée par Sprinzak *et al.* Cette méthode d'association définit une mesure de l'interaction entre deux domaines, en calculant la fraction de paires de protéines en interaction, par rapport à toutes les paires de protéines avec ces domaines [Sprinzak et Margalit, 2001]. Par conséquent, ceci peut attribuer des scores d'association élevés à des paires de domaines apparaissant avec une fréquence faible. Kim *et al.* ont amélioré cette méthode d'association en considérant le nombre de domaines de chaque protéine [Kim *et al.*, 2002]. Une approche intégrative a également été proposée par Ng *et al.* pour prédire des interactions domaine-domaine en intégrant des informations provenant de trois sources différentes [Ng *et al.*, 2003a] : des interactions protéine-protéine identifées expérimentalement, des complexes de protéines et des séquences de protéines de fusion (*Rosetta Stone Sequence*). Ils ont regroupé leurs résultats dans la base de données InterDom [Ng *et al.*, 2003b]. Les scores d'interactions des paires de domaines pour chaque source de données sont obtenus de la même façon que pour la méthode d'association, en considérant la fréquence de chaque domaine dans les paires de protéines.

Ces différentes approches considèrent seulement des paires de domaines, et supposent que les interactions entre les paires de domaines sont indépendantes les unes des autres. Pourtant, certaines protéines contiennent plusieurs domaines, et les interactions protéine-protéine peuvent être le résultat de multiples paires de domaines, ou de groupes de domaines. Han *et al.* ont ainsi proposé une méthode basée sur la combinaison des domaines et les paires de combinaisons de domaines [Han *et al.*, 2004a]. Les protéines sont alors décomposées selon les différentes combinaisons possibles de leurs domaines (voir Figure 3.7). Par contre, toutes ces approches basées sur la méthode d'association souffrent d'une limitation générale qui est de ne pas tenir compte de l'accessibilité des différents domaines, et donc de la faisabilité des interactions.

Wojcik *et al.* ont, quant à eux, proposé une approche basée sur les graphes [Wojcik et Schächter, 2001] appelée $IDPP$ pour *Interacting Domain Pair Profil*. La méthode utilise à la fois l'homologie de séquence et le clustering basé sur des profils d'interaction, ainsi que l'information apportée par les domaines. Les auteurs ont montré qu'ils obtiennent des prédictions d'interactions protéine-protéine meilleures qu'en utilisant seulement l'information de la séquence.

3.3.4 Méthodes basées sur la structure

D'autres méthodes utilisent des informations plus proches de la structure tridimensionnelle de la protéine, et cherchent d'abord à identifier les sites d'interaction. Ceci peut se faire par la reconnaissance de motifs de résidus [Kini et Evans, 1996], ou en utilisant les propriétés de la topologie de l'interface, de la surface accessible au solvant ou de l'hydrophobicité [Jones et Thornton, 1997].

3.3.5 Méthodes d'apprentissage

Différentes méthodes de classification avec apprentissage ont été utilisées pour prédire des interactions entre des protéines, ou entre des domaines. L'idée est d'apprendre les caractéristiques des paires de protéines en interaction, afin de prédire pour deux protéines quelconques si elles interagissent ou non, ou quelle est la probabilité qu'elles interagissent. On peut citer par exemple les méthodes des noyaux [Yamanishi *et al.*, 2004], [Ben-Hur et Noble, 2005], [Geurts *et al.*, 2007], les méthodes basées sur les forêts aléatoires (random forest decision) [Qi *et al.*, 2005], [Chen et Liu, 2005], les machines à vecteurs de support (SVM) [Bradford et Westhead, 2005], les réseaux de neurones [Koike et Takagi, 2004], [Fariselli *et al.*, 2002], ou encore un grand nombre de méthodes statistiques basées en général sur des modèles bayésiens [Deng *et al.*, 2002],[Jansen *et al.*, 2003b], [Lee *et al.*, 2006].

Pour finir, une méthode a récemment été développée pour prédire des relations fonctionnelles à partir de l'usage des codons [Najafabadi et Salavati, 2008], c'est-à-dire basée simplement sur la séquence (voir page 42).

Nous avons parcouru ici un ensemble de méthodes de prédiction permettant d'identifier des relations fonctionnelles entre des gènes ou des protéines, ces relations pouvant être des interactions physiques dans certains cas. Par la suite, nous nous intéresserons plus particulièrement à la première méthode présentée, la méthode par interologue, qui consiste à transférer des interactions d'un organisme vers un autre en utilisant des relations d'orthologie entre les protéines. La fin de cette étude bibliographique est consacrée aux méthodes d'intégration de données.

Chapitre 4

Intégration des données

"Les gènes, comme les diamants, sont éternels,
mais pas tout à fait de la même façon que ces derniers."

Richard Dawkins,
Le gène égoïste, 1976

Nous présentons d'abord les questions biologiques principales qui sont posées et amènent les chercheurs à intégrer différents types d'information. Les stratégies adoptées par certaines méthodes d'intégration sont ensuite exposées.

4.1 Présentation des questions biologiques

Dans le domaine de la recherche biomédicale, un des principaux objectifs est d'identifier des cibles thérapeutiques, c'est-à-dire des gènes qui sont impliqués dans certaines maladies, en particulier les cancers, directement ou par l'intermédiaire de leurs produits (ARNs, protéines). La connaissance de ces cibles permet de diriger entre autres la conception de médicaments. Jeong *et al.*, ainsi que Yu *et al.*, ont par exemple proposé des critères topologiques pour prédire les gènes essentiels, c'est-à-dire des gènes dont l'absence est létale, et par conséquent importants pour l'organisme [Jeong *et al.*, 2003], [Yu *et al.*, 2007].

Plus généralement, l'objectif est de comprendre les mécanismes cellulaires de régulation, par exemple les mécanismes de réponse aux stress oxydants et aux métaux lourds. Pour cela, la description et la compréhension se font principalement à deux niveaux : global et local.

À l'échelle globale, le but est d'obtenir une description complète des systèmes génétiques de contrôles cellulaires. Ceci se traduit notamment par deux problèmes : l'annotation des gènes et la modélisation des réseaux biologiques. Annoter un gène consiste à décrire ses fonctions. Il est souvent nécessaire de prédire les fonctions quand celles-ci sont inconnues. Pour cela, des méthodes ont été développées qui utilisent les relations entre les gènes et les annotations déjà connues, de manière à prédire les fonctions encore inconnues de certains gènes. Brun *et al.* ont notamment proposé une méthode de

prédiction de fonction basée sur les réseaux d'interactions protéine-protéine et les voisinages des protéines dans ces réseaux [Brun *et al.*, 2003]. De plus, McDermott *et al.* ont proposé une méthode de pondération des voisins (*neighborhood weighting method*) [McDermott *et al.*, 2005].

La modélisation des réseaux biologiques concerne non plus les gènes séparément mais les relations entre eux. Ceci consiste notamment à décrire les structures générales des réseaux biologiques, et à modéliser les relations entre les acteurs biologiques (gène, ARN, protéine, métabolite), par exemple sous la forme de voies moléculaires (métaboliques ou signalisation). Albert *et al.*, ainsi que Jeong *et al.*, ont par exemple montré que les réseaux d'interactions protéine-protéine semblaient avoir une topologie appelée scale-free [Albert *et al.*, 2000], [Jeong *et al.*, 2001], c'est-à-dire que la plupart des protéines sont en interaction avec peu d'autres protéines, alors qu'un petit nombre d'entre elles possèdent énormément de partenaires d'interaction. De plus, Wang *et al.* ont posé la question de la pertinence des réseaux actuels, dans la mesure où certaines méthodes expérimentales détectent des complexes entre plusieurs protéines qui sont ensuite traduits en plusieurs interactions binaires [Wang et Zhang, 2007]. Concernant les voies de signalisation ou les voies métaboliques, différentes méthodes ont été développées pour extraire cette information à partir de réseaux d'interactions protéine-protéine [Ideker *et al.*, 2002], [Segal *et al.*, 2003], [Calvano *et al.*, 2005], [Scott *et al.*, 2006].

À l'échelle locale, l'accent est mis sur la description détaillée d'un mécanisme d'intérêt, centrée autour d'un ensemble de protéines directement impliquées dans ce mécanisme. Pour répondre à ce problème, différentes stratégies sont utilisées. D'un côté les études classiques de biochimie permettent d'obtenir des résultats détaillés, mais sont coûteuses et nécessitent beaucoup de temps pour être mises en place de manière effective. D'un autre côté, les approches dites à haut-débit permettent au contraire d'avoir des résultats étendus rapidement, mais produisent un grand nombre d'erreurs. Par conséquent, les méthodes qui combinent ces deux aspects sont utiles afin de compléter la connaissance déjà acquise de certains réseaux biologiques clefs. Pour cela, il faut combiner la génétique classique et les techniques de biologie cellulaire avec l'analyse des données de génomique fonctionnelle obtenues par des approches à haut-débit. L'objectif est notamment d'identifier les gènes clefs et les voies moléculaires impliqués dans la régulation cellulaire, le développement ou les maladies en se focalisant sur certains aspects. Calvano *et al.* ont par exemple étudié les leucocytes du sang chez des patients humains ayant reçu un stimulus inflammatoire [Calvano *et al.*, 2005]. Par ailleurs, dans le domaine des métastases en cancérologie, Jonsson *et al.* ont concentré leur étude autour de quelques protéines à fort potentiel métastatique chez le rat [Jonsson *et al.*, 2006].

Il est évident que toutes ces questions sont très liées les unes avec les autres. Ainsi, les différentes études répondent à plusieurs questions biologiques en même temps et ces problématiques progressent en parallèle. Pour répondre à l'ensemble de ces questions, des méthodes d'intégration de différentes données ont été développées. Dans sa revue, Troyanskaya montre en particulier comment les données transcriptome, qui ont fait partie, après le séquençage, des résultats des premières approches à haut-débit, sont maintenant complétées par d'autres données biologiques [Troyanskaya, 2005]. Combiner

différents types de données a entre autres les trois objectifs suivants : combiner les forces, combiner les spécificités et tirer profit des relations qui existent entre ces données. Nous allons développer ces trois objectifs dans la partie suivante.

4.2 Présentation des méthodes d'intégration

Combiner les forces des différentes données permet d'obtenir des résultats plus précis et plus riches.

Une stratégie utilisée est de considérer l'intersection de plusieurs réseaux d'interactions. Tong *et al.* ont ainsi proposé une approche systématique pour identifier les réseaux d'interactions protéine-protéine dans lesquels plusieurs domaines de reconnaissance de peptide sont impliqués [Tong *et al.*, 2002]. Cette stratégie a l'avantage de réduire le nombre de faux positifs, mais a l'inconvénient en contre-partie de réduire la sensibilité. C'est en effet le jeu de données qui a la plus faible sensibilité qui limite la sensibilité de l'analyse entière. Comme les études à grande échelle publiées sont loin d'être complètes même pour les organismes modèles [Hart *et al.*, 2006], cette limitation est un inconvénient important. D'autres études ont permis d'adapter cette approche en augmentant la sensibilité. Marcotte *et al.* ont par exemple prédit des interactions protéine-protéine chez la levure en combinant différents jeux de données à l'aide d'une heuristique. Ainsi, ils considèrent les interactions de qualité élevée ou prédites par au moins deux méthodes [Marcotte *et al.*, 1999b].

À l'inverse, l'union des jeux de données peut être prise en compte, augmentant alors la sensibilité mais augmentant aussi le nombre de faux positifs.

Gerstein *et al.* ont essayé quant à eux de se positionner plus finement que dans les cas extrêmes où l'intersection ou l'union sont considérées, c'est-à-dire dans une position intermédiaire. Pour cela, ils ont pris en compte les taux relatifs de faux positifs et de faux négatifs des différents jeux de données [Gerstein *et al.*, 2002]. Dans une autre étude, Schwikowski *et al.* ont assigné des fonctions potentielles en se basant sur le nombre d'interactions qu'une protéine inconnue avait avec des protéines de différentes catégories fonctionnelles [Schwikowski *et al.*, 2000]. Par ailleurs, lors de leur étude sur la qualité des interactions protéine-protéine, von Mering *et al.* ont constaté que les réseaux constitués des interactions mises en évidence par plusieurs méthodes obtenaient une meilleure précision que les réseaux construits séparément par chacune des méthodes expérimentales [von Mering *et al.*, 2002].

Combiner les spécificités des différents jeux de données permet d'obtenir des résultats plus complets, chaque type de données apportant des informations complémentaires. Les données d'interaction protéine-protéine mettent en évidence l'ensemble des interactions possibles entre des protéines d'un organisme donné. Mais ces interactions peuvent être plus ou moins stables dans le temps, comme dans le cas de complexes, ou être de nature plus transitoire et ne pas avoir lieu en même temps. Les données transcriptome peuvent compléter ces informations en apportant un aspect dynamique. Ainsi, Komurov *et al.* ont mis en évidence des propriétés statiques et dynamiques de l'architecture modulaire du réseau d'interactions protéine-protéine chez la levure [Komurov et White, 2007]. D'autre

part, Myers *et al.* ont utilisé des données fonctionnelles pour rendre compte du contexte biologique des données d'interactions [Myers et Troyanskaya, 2007].

En plus de combiner les forces et les spécificités des différents jeux de données, les relations entre ces données sont utilisées en général pour décrire d'abord, puis corroborer et enfin prédire.

Différentes études ont mis en évidence des relations entre plusieurs types de données, par exemple entre interactome et transcriptome, ou entre la topologie d'un réseau et les caractéristiques fonctionnelles de ses éléments. Grigoriev *et al.* ont d'abord montré l'existence d'une relation entre l'expression des gènes et les interactions protéine-protéine au niveau du protéome [Grigoriev, 2001]. Ainsi, les paires de protéines codeés par des gènes co-exprimés, c'est-à-dire avec des profils d'expression proches, interagissent plus fréquemment que des paires de protéines aléatoires. De même, Ge *et al.* ont souligné la corrélation entre le transcriptome et l'interactome chez la levure [Ge *et al.*, 2001]. Pour cela, ils ont défini des classes de gènes montrant un profil d'expression commun. Ensuite, ils ont comparé les interactions entre des protéines codées par des gènes appartenant à une même classe, avec les interactions codées par des gènes appartenant à des classes différentes. Si ce résultat paraissait cohérent avec l'idée que des protéines d'un même complexe ou d'une même voie de signalisation étaient co-régulées, les auteurs ont toutefois souligné l'existence d'exceptions déjà décrites. En effet, au cours du cycle cellulaire, les gènes qui codent les CDK (*cyclin-dependant kinases*) ont un profil d'expression constant, c'est-à-dire qu'ils ne sont pas régulées. On parle alors d'expression constitutive. En revanche, les gènes qui codent leurs sous-unités de régulation, les cyclines, sont largement régulés. Cette corrélation entre l'interactome et le transcriptome est donc une propriété simplificatrice. Elle peut être utilisée pour améliorer les hypothèses émergeant individuellement de ces données d'interactome ou de transcriptome. Par ailleurs, Wach *et al.* ont étudié la relation entre interactome et transcriptome dans le contexte du cancer du poumon [Wachi *et al.*, 2005]. À cette occasion, ils ont mis en évidence que les gènes différentiellement exprimés entre des tissus sains et malades étaient fortement connectés dans le réseau d'interactions protéine-protéine.

D'autre part, la relation entre la topologie d'un réseau biologique et les propriétés fonctionnelles de ses éléments est une question importante. Ainsi, Jeong *et al.* ont montré que dans un réseau d'interactions protéine-protéine, les protéines fortement connectées, appelées *hubs*, étaient souvent essentielles à la survie de l'organisme [Jeong *et al.*, 2001]. La raison de cette corrélation est toujours à l'étude [Zotenko *et al.*, 2008]. Néanmoins, il a été montré récemment que cette corrélation n'apparaît pas dans les réseaux d'interactions obtenus par l'approche double-hybride [Yu *et al.*, 2008].

Par la suite, Fraser *et al.* ont mis en évidence une corrélation négative entre le taux d'évolution des protéines bien conservées, et leur connectivité au sein d'un réseau d'interactions protéine-protéine [Fraser *et al.*, 2002]. Ces relations ont ensuite été enrichies avec des données transcriptome, afin d'y ajouter une notion de dynamique ou de contexte biologique. Ainsi, Han *et al.* ont étudié comment les protéines à fort degrés (*hubs*) pouvaient contribuer aux propriétés cellulaires des interactions protéine-protéine régulées dynamiquement à la fois dans le temps et dans l'espace. Pour cela, ils ont considéré les

hubs et leur co-expression avec l'ensemble de leurs voisins, c'est-à-dire les protéines avec lesquelles ils sont en interaction [Han *et al.*, 2004b]. Ils ont ainsi défini deux classes de *hubs* : les *party hubs*, qui interagissent simultanément avec tous leurs partenaires ; les *date hubs*, qui se lient avec des partenaires différents, à différents instants ou en différents endroits. Ils ont alors proposé un modèle d'organisation modulaire dans lequel les *date hubs* organisent le protéome en connectant entre eux les processus biologiques, ou modules, alors que les *party hubs* fonctionnent à l'intérieur des modules. Ce modèle a été rejeté par Batada *et al.* en utilisant des interactions mises en évidence entre temps et un filtrage sur la qualité des ces dernières [Batada *et al.*, 2006]. Cependant, Bertin *et al.* ont confirmé leurs résultats, c'est-à-dire l'organisation modulaire mise en place par les *date* et *party hubs*, en analysant les données récentes des interactions protéine-protéine chez la levure [Bertin *et al.*, 2007].

D'autres études ont été faites pour voir dans quelle mesure ces relations étaient vraies selon les réseaux considérés et quelles pouvaient être les éventuelles différences. Ainsi, Jansen *et al.* ont étudié la relation entre les interactions protéine-protéine et les niveaux d'expressions des gènes correspondants, dans le cadre de complexes bien définis pour lesquels des interactions ont été montrées entre leurs sous-unités. Ils ont d'abord confirmé que les sous-unités d'un même complexe étaient significativement co-exprimées. Toutefois, ils ont précisé ce résultat en montrant que les complexes permanents, comme le ribosome ou le protéasome, avaient en général une relation particulièrement forte avec le niveau d'expression, contrairement aux complexes transitoires [Jansen *et al.*, 2002]. Enfin, ils ont établi que les interactions mises en évidence par double-hybride avaient une relation assez faible avec le niveau d'expression, de la même façon que les complexes transitoires.

Par ailleurs, Yu *et al.* ont étudié la structure dynamique de deux sortes de réseaux biologiques (réseau d'interactions protéine-protéine et réseau de régulation) en y associant des données transcriptome [Yu *et al.*, 2007]. Ils ont défini à cette occasion les *bottlenecks* comme les protéines d'un réseau biologique qui sont traversées par un grand nombre de plus courts chemins. En d'autres termes, ces protéines sont des points de passage obligatoire pour la communication entre un grand nombre de protéines. Les auteurs ont montré que ces protéines étaient largement susceptibles d'être des protéines essentielles. Dans le cas des réseaux de régulation, ce critère était d'après eux un meilleur outil de prédiction pour l'essentialité que celui basé sur le fort degré (*hubs*) [Jeong *et al.*, 2001], [Yu *et al.*, 2004a]. De plus, ils ont montré que ces protéines *bottlenecks* correspondaient à la composante dynamique des réseaux biologiques. Néanmoins, les auteurs ont pu mettre en évidence des différences entre les réseaux d'interactions protéine-protéine et les réseaux de régulation. En particulier, le degré des protéines reste le meilleur critère de prédiction de l'essentialité pour les réseaux d'interactions protéine-protéine. Les auteurs ont expliqué ces différences par le fait que les réseaux de régulation sont orientés et supportent par conséquent un flot d'information implicite, ce qui les rend assez similaires à un système de transport. Par contre, les réseaux d'interactions protéine-protéine ne sont pas orientés et il contiennent moins clairement un flot d'information.

Après la phase descriptive, les relations entre les jeux de données ont été supposées

et ont servi à corroborer de nouveaux résultats, c'est-à-dire à renforcer, pondérer ou qualifier des interactions. Von Mering *et al.* ont notamment comparé des interactions protéine-protéine en analysant les relations fonctionnelles des protéines en interaction [von Mering *et al.*, 2002]. Les auteurs ont en effet proposé de mesurer la qualité des interactions protéine-protéine en calculant le degré auquel les protéines en interaction étaient annotées avec la même catégorie fonctionnelle. Ils ont ainsi suggéré que les interactions entre des protéines annotées différemment étaient majoritairement des faux positifs, c'est-à-dire des interactions n'ayant pas lieu en réalité. Par conséquent, ils ont franchi un pas important par rapport à la position de Ge *et al.*En effet, ces derniers ont utilisés les relations entre les jeux de données pour renforcer des résultats, quand par exemple des protéines en interactions dont les gènes correspondants sont également co-exprimés, mais ils ne considéraient pas ces relations comme nécessaires. Ainsi, le fait que les gènes ne soient pas co-exprimés n'affaiblit pas pour autant l'information obtenue sur l'interaction entre les protéines. De même, Kemmeren *et al.* ont utilisé un jeu de données de nature différente afin d'évaluer la qualité des interactions protéine-protéine chez la levure [Kemmeren *et al.*, 2002]. Ils ont exploité pour leur part des données transcriptome en se basant sur cette même hypothèse de corrélation entre les interactions et la co-expression. En conséquence, ils ont considéré que la confiance qu'ils avaient dans un jeu de données d'interactions pouvait être augmentée en se fondant sur les protéines en interaction dont les gènes correspondants étaient co-exprimés. En outre, Poyatos *et al.* ont utilisé les relations phylogénétiques pour vérifier la décomposition en modules de réseaux d'interactions protéine-protéine [Poyatos et Hurst, 2004].

Finalement, la tâche la plus compliquée est la phase prédictive. La principale approche adoptée consiste à construire des modèles probabilistes permettant de combiner des données hétérogènes de manière générale. Ces méthodes probabilistes fournissent par définition des probabilités qui peuvent être utilisées pour filtrer les résultats à un niveau souhaité de sensibilité ou de spécificité. Troyanskaya *et al.* ont notamment utilisé une approche bayésienne afin de construire un réseau d'interactions fonctionnelles qu'ils ont ensuite exploité pour prédire la fonction des gènes inconnus d'après leur réseau [Troyanskaya *et al.*, 2003]. De plus, Jansen *et al.* ont développé un modèle bayésien pour prédire cette fois des interactions protéine-protéine chez la levure à l'échelle du génome [Jansen *et al.*, 2003b].

Nous avons exposé ici les questions principales dirigeant les recherches actuelles sur l'intégration des données, ainsi que quelques travaux majeurs.

Deuxième partie

Démarche

Chapitre 1

Caractérisation des classes de gènes régulés en réponse aux stress oxydants et aux métaux lourds

"Mais l'homme ne se borne pas à voir ; il pense et veut connaître la signification des phénomènes dont l'observation lui a révélé l'existence."

Claude Bernard,
Introduction à la médecine expérimentale, 1966

Dans ce chapitre, nous avons souhaité décrire la régulation de la transcription en réponse à différents stress environnementaux, c'est-à-dire voir ce qui se passe, principalement à l'échelle des gènes et des protéines correspondantes, lorsque la cellule est soumise à un stress déclenché par un agent oxydant ou un métal. Pour cela, nous avons notamment voulu caractériser la réponse transcriptionnelle de manière globale, c'est-à-dire savoir combien de gènes et lesquels étaient induits ou réprimés. Le but était ensuite d'identifier les gènes co-exprimés, c'est-à-dire qui sont exprimés de manière similaire dans différentes conditions. Pour cela, nous avons recherché les classes de gènes dont les comportements étaient proches en termes de réponse transcriptionnelle. Finalement, notre objectif était d'expliquer pourquoi certaines classes de gènes étaient induites alors que d'autres étaient réprimées. Pour cela, nous avons voulu caractériser les sous-ensembles de protéines correspondantes par des signatures, telles que la teneur en soufre ou l'hydrophobicité, afin de comparer ces classes de protéines.

1.1 Présentation du dispositif expérimental

Pour étudier la régulation de la transcription en réponse aux stress oxydants et métalliques, nous avons choisi des agents induisant un stress, un organisme d'étude, ainsi qu'un ensemble de protocoles expérimentaux afin de conduire des expériences appropriées à la problématique traitée.

1.1.1 Choix des agents inducteurs de stress

Dans cette étude, nous avons étudié en particulier les réponses cellulaires à deux types de stress, les stress oxydants et les stress métalliques.

Le stress oxydant est notamment présent chez les organismes photosynthétiques. La photosynthèse est le processus bioénergétique qui permet aux plantes de synthétiser leur matière organique en exploitant l'énergie solaire. Plus précisément, c'est la fabrication de matière carbonée organique, à partir d'eau et de carbone minéral (CO_2) en présence de lumière. Ce processus se fait par la réduction des matières inorganiques et la libération de molécules de dioxygène. L'utilisation de l'énergie solaire par les organismes photosynthétiques a donc comme conséquence le renouvellement de l'oxygène de l'atmosphère. Par ailleurs, la respiration végétale est le processus de respiration qui a lieu dans une plante. Ce processus biologique se traduit par une consommation de dioxygène et un rejet de carbone minéral (CO_2), comme chez les animaux. Lors de ces deux processus biologiques, des espèces réactives de l'oxygène, appelées ROS (*Reactive Oxygen Species*), sont produites comme les ions oxygènes, les radicaux libres ou les peroxydes. Ces espèces chimiques sont toxiques pour les organismes vivants car elles entraînent l'oxydation de certains composants de la cellule comme par exemple les protéines, les lipides, ou l'ADN. Pour cette raison, les organismes photosynthétiques sont souvent soumis à des stress oxydants.

Le stress métallique, pour sa part, est induit par les métaux. Les métaux peuvent être classés en deux groupes principaux. Les métaux dits essentiels participent au métabolisme des organismes vivants. Il s'agit par exemple du fer ou du zinc. Ces métaux sont nécessaires au bon fonctionnement de l'organisme, mais s'ils ne sont pas présents dans les bonnes quantités, ils peuvent entraîner un stress par excès ou par carence. L'autre catégorie de métaux, les métaux dits toxiques, ne participent pas au métabolisme des organismes vivants. Il s'agit entre autres du cadmium, du mercure ou du plomb. Ces métaux lourds entraînent un stress car ils peuvent par exemple réagir avec des protéines, inactiver des enzymes. Dans certains cas, ils remplacent les éléments métalliques, comme le fer ou le zinc, qui font habituellement partie des métallo-enzymes. Il a été montré que ces stress métalliques conduisent souvent à un stress oxydant [Stohs et Bagchi, 1995]. Par conséquent, les stress oxydants et métalliques sont liés.

Pour étudier les mécanismes de réponse aux stress oxydants et aux métaux lourds, nous avons choisi d'analyser en particulier les réponses cellulaires aux stress induits par la présence de cadmium (Cd), la présence de péroxyde d'hydrogène (H_2O_2), la carence en fer ($-$Fe), l'excès de fer ($+$Fe) et l'excès de zinc ($+$Zn).

1.1.2 Choix de l'organisme d'étude

Le choix de l'organisme d'étude a quant a lui été guidé par différentes raisons. De manière générale, les cyanobactéries présentent pour cette étude trois avantages principaux.

Tout d'abord, ce sont les organismes photosynthétiques les plus abondants sur Terre et ce sont des modèles d'autant plus attractifs qu'elles réalisent les deux processus biologiques de la respiration et de la photosynthèse dans le même système membranaire [Peschek, 1996].

De plus, les cyanobactéries partagent un grand nombre de gènes avec les plantes [Martin *et al.*, 2002], ce qui est en accord avec le scénario évolutif selon lequel les chloroplastes des plantes proviendraient de l'endosymbiose de cyanobactéries ancestrales [Gray, 1993]. Par conséquent, une meilleure compréhension des mécanismes de réponse aux stress oxydants et métalliques chez les cyanobactéries permettra de mieux comprendre comment les plantes réagissent à ces mêmes stress.

Enfin, les cyanobactéries permettent des applications intéressantes et prometteuses. Elles peuvent notamment être utilisées comme rapporteurs biologiques [Bachmann, 2003]. Dans ce cas, elles sont modifiées génétiquement de manière à ce que des gènes rapporteurs aident à déterminer la bio-disponibilité de certaines espèces chimiques et leurs effets sur les organismes vivants. Elles peuvent également être utilisées dans le cadre de la bioremédiation [Gong *et al.*, 2005], c'est-à-dire l'ensemble des techniques dont l'objectif est de lutter contre les pollutions en augmentant la biodégradation ou la biotransformation. Les cyanobactéries pourraient en outre être utilisées dans le domaine de la bio-énergétique pour produire de l'hydrogène. Cette production biologique d'hydrogène, ainsi que d'autres méthodes de production actuellement à l'étude telles que l'électrolyse ou la dissociation de l'eau, pourraient compléter les méthodes de production basées sur les sources d'énergie fossile.

Parmi les cyanobactéries, nous nous sommes intéressés à *Synechocystis* PCC6803, appelé *Synechocystis* dans la suite, pour trois raisons principales. Cet organisme modèle a un petit génome (3 500 gènes environ), ce qui facilite à la fois les manipulations génétiques et les analyses. De plus, le génome de cette cyanobactérie est entièrement séquencé depuis 1996 [Kaneko *et al.*, 1996]. Son génome est décrit dans le base de données Cyanobase [Nakamura *et al.*, 1998], [Nakamura *et al.*, 2000]. Enfin, *Synechocystis* est facilement manipulable avec les outils développés par le laboratoire [Poncelet *et al.*, 1998], [Mazouni *et al.*, 2003], [Mazouni *et al.*, 2004].

1.1.3 Choix des procédures expérimentales

Pour étudier la régulation de la transcription des gènes en réponse aux stress étudiés, des puces à ADN ont été utilisées (voir page 29). Celles-ci permettent de comparer les niveaux d'expression relatifs, pour chaque gène, d'un échantillon cellulaire soumis au stress étudié par rapport à ceux d'un échantillon de contrôle, c'est-à-dire n'ayant pas subi ce stress. Ces procédures expérimentales ont été réalisées principalement par Lætitia Houot. Nous en présentons rapidement les points clefs nécessaires à la compréhension

des analyses.

En général, les cyanobactéries sont cultivées dans un milieu liquide nutritif appelé BG11. Ce milieu minéral contient $17\mu M$ de fer et $0{,}77\mu M$ de zinc. Nous avons considéré ces conditions de culture comme la référence pour des conditions dites normales. Nous avons choisi d'étudier notamment deux types de stress subis par les organismes vivants : les stress à caractère permanent et les stress induits par des variations brutales de l'environnement.

Pour mimer les stress permanents, les cellules ont été mises de manière continue en présence d'un agent inducteur de stress. Ainsi, des cellules en croissance ont été exposées à du sulfate de cadmium ($CdSO_4$, $50\mu M$) ou à du péroxyde d'hydrogène (H_2O_2, 3mM) pendant des périodes de temps de plus en plus longues. Ceci nous a permis d'obtenir, pour chaque agent, un ensemble ordonné de points de mesure que l'on appelle une cinétique (15, 30, 60, 75, 90 180, 300, 360 et 960 minutes pour le Cd, voir la Table 1.1 et 15, 30, 180, 300 et 420 minutes pour H_2O_2, voir la Table 1.2). La même stratégie a été menée dans le cas de stress induits par excès de fer ($(NH_4)FeH_2C_6H_5O_7$, 17mM, voir la Table 1.2) et excès de zinc ($ZnSO_4$, $776\mu M$, voir la Table 1.2). Il est difficile de parler de cinétique dans ce cas puisque deux points de mesures ont été réalisés seulement pour chaque agent (240 et 360 minutes pour +Fe, 30 et 240 minutes pour +Zn, voir la Table 1.2).

Dans une deuxième approche, des changements drastiques dans la disponibilité des nutriments ont été mis en place. Ainsi, dans le cas du stress induit par la carence en fer ($-$Fe), les cellules ont d'abord été immergées dans un milieu contenant $2\mu M$ de fer ($(NH_4)FeH_2C_6H_5O_7$), puis dans un milieu sans fer. Cette procédure est dénotée 2-0. De même, la procédure dénotée 1-0 correspond à une immersion dans un milieu contenant $1\mu M$ de fer, puis dans un milieu sans fer (voir la Table 1.2).

Pour chaque point de mesure, plusieurs réplicats ont été effectués. Dans la majorité des cas, un seul *dye-swap* a été réalisé. Un *dye-swap* est constitué de deux réplicats pour lesquels les marqueurs (Cy_3 et Cy_5, voir page 29) ont été échangés entre les mesures respectives de manière à s'affranchir des biais pouvant exister dans l'assignation d'un canal donné à un échantillon biologique (échantillon traité ou échantillon de contrôle). Dans certains cas, la manipulation a été faite deux fois, produisant alors quatre réplicats en tout pour le point de mesure considéré (voir la Table 1.1).

Après avoir mené ces expériences, nous avons voulu exploiter les résultats expérimentaux afin de caractériser les réponses cellulaires de *Synechocystis* aux stress oxydants et métalliques présentés.

1.2 Caractérisation de la réponse cellulaire

Nous avons d'abord voulu mener une analyse préliminaire de manière à dégager les principales tendances des réponses transcriptionnelles étudiées. Pour cela, nous avons normalisé les données brutes et utilisé une méthode heuristique pour déterminer si un

gène donné était induit ou réprimé à un temps donné.

Cette analyse nous ayant permis de mettre en évidence des phases de réponse, nous avons développé dans un second temps une analyse statistique, afin d'identifier les gènes mettant en évidence un comportement différent au cours des phases de réponse. Finalement, nous avons voulu comprendre les mécanismes biologiques sous-jacents.

1.2.1 Normalisation des données

Comme expliqué dans l'étude bibliographique (voir page 29), les puces sont scannées par un laser en utilisant différentes longueurs d'onde de manière à obtenir les intensités numériques de chaque spot. Ainsi, une mesure relative à l'intensité globale d'hybridation est obtenue pour chaque élément sur la puce. Dans notre cas, les puces à ADN ont été scannées avec le logiciel GenePix ($GenePix^{TM}$ Pro 4.0). Or, ces données de transcriptome sont bruitées et biaisées (voir page 31).

Dans un premier temps, il était donc nécessaire de normaliser les données brutes. Nous avons alors choisi d'appliquer la méthode de normalisation la plus classiquement utilisée, à savoir le traitement du bruit de fond, puis le traitement du biais par rapport à l'intensité totale. Pour cela, nous avons tout d'abord soustrait le bruit de fond pour obtenir l'intensité du signal pour chaque spot. Nous avons ensuite utilisé le logiciel TIGR ExpressConverter (version 1.7) [Saeed *et al.*, 2003] pour appliquer la normalisation lowess [Cleveland et Devlin, 1988] en nous basant sur la fonction *locfit* du logiciel TIGR Midas (version 2.19). Nous avons réglé le paramètre de lissage (smooth) à 0,33 comme cela est recommandé [Quackenbush, 2002] (voir page 32 pour le principe de cette normalisation). Les valeurs normalisées de chacun des deux canaux (Cy_3 et Cy_5) ont ensuite été combinées pour obtenir un ratio (voir Équation 1.1) ou un log-ratio (voir Équation 1.2).

$$ratio = \frac{Cy_3}{Cy_5} \tag{1.1}$$

$$ratio = log_2(ratio) = log_2(\frac{Cy_3}{Cy_5}) \tag{1.2}$$

Lorsque cela était nécessaire, nous avons regroupé les valeurs des réplicats, par exemple pour comparer les différents points de mesure. Ainsi, pour chaque point de mesure, les différentes valeurs ont été moyennées de manière à obtenir un unique ratio par gène pour un temps donné. Pour cela, nous avons calculé la moyenne arithmétique des log-ratios [Quackenbush, 2002], ce qui correspond à la moyenne géométrique des ratios. Après que les données brutes ont été normalisées de cette façon, nous avons mené une première analyse globale.

1.2.2 Analyse préliminaire des résultats

Une heuristique a tout d'abord été utilisée pour déterminer les tendances majeures des réponses transcriptionnelles des cellules soumises aux différents stress étudiés. Pour cela, nous avons sélectionné deux groupes de gènes : les gènes induits étaient ceux dont

le ratio était supérieur à 1,9 ; les gènes réprimés étaient ceux dont le ratio était inférieur à 0,52 ($= 1/1,9$).

Pour la réponse transcriptionnelle suite à un stress Cd, nous avons identifié deux phases. La première phase s'étale du point de départ jusqu'à 60 minutes (voir la Table 1.1). Elle est plutôt modérée dans la mesure où seulement 210 gènes sont différentiellement exprimés. Ces gènes sont majoritairement induits (63%). La seconde phase se produit entre 90 et 360 minutes. Elle est massive puisque 1 315 gènes sont différentiellement exprimés (37% du génome). Ces gènes sont répartis entre les induits et les réprimés de manière assez équilibrée (692 réprimés et 623 induits).

Concernant la réponse au stress induit par H_2O_2, elle apparaissait plus rapide et plus courte (voir la Table 1.2). Nous avons également identifié deux phases principales. La première phase comprend les deux premiers temps de la cinétique (15 et 30 minutes). Elle est massive car 1 459 gènes sont différentiellement exprimés. Ces gènes sont répartis entre les induits et les réprimés de manière assez équilibrée (781 réprimés et 678 induits). La seconde phase se situe à partir de 180 minutes. Seulement 446 gènes sont encore différentiellement exprimés au cours de cette phase (177 réprimés et 269 induits). Nous avons qualifiée cette phase de tardive dans la mesure où il s'agit d'une période pendant laquelle la plupart des gènes ayant répondu rapidement et fortement retrouvent un niveau d'expression normal, c'est-à-dire à peu près similaire à celui des gènes des cellules non traitées.

Les deux autres expériences réalisées dans le cas d'un stress permanent, à savoir l'excès de Fe et l'excès de Zn, n'ont pas permis de mettre en évidence des phases de réponse puisque seulement deux temps différents ont été considérés pour chaque stress. Dans le cas de l'excès de Zn d'abord, 28% des gènes sont différentiellement exprimés (406 réprimés et 429 induits). Très peu de gènes sont régulés lors du premier temps (60 gènes à 30 min). La réponse est plus importante lors du second temps (466 gènes régulés à 240 min). Ensuite, dans le cas de l'excès de Fe, 10% des gènes sont différentiellement exprimés (100 réprimés et 201 induits). Dans ce cas-ci, les deux temps (240 et 360 min) sont relativement similaires en termes de nombre de gènes induits et réprimés (voir Table 1.2).

Enfin, dans le cas du stress ponctuel mimé par la carence en Fe, il ne peut y avoir de phase de réponse puisqu'un temps unique est considéré. Certes une cinétique pourrait être construite en considérant les concentrations utilisées et non les temps, mais là encore, seules deux concentrations ont été considérées, ne permettant pas l'identification de phases. Dans le cas de ce stress, 15% des gènes sont différentiellement exprimés (226 réprimés et 220 induits). Pour les deux concentrations, les nombres de gènes induits et réprimés sont proches (voir Table 1.2).

Après avoir recherché les gènes régulés et identifié ces phases de réponse par une méthode heuristique, nous avons voulu vérifier la pertinence de ces résultats par une méthode statistique plus appropriée.

1.2.3 Identification des gènes globalement régulés

Afin d'identifier les gènes globalement régulés, c'est-à-dire induits ou réprimés au cours des réponses transcriptionnelles lors des différents stress étudiés, nous avons choisi d'utiliser un modèle linéaire. Pour cela, nous avons considéré les différents réplicats de tous les points de mesure de façon séparée, c'est-à-dire que nous n'avons pas calculé la moyenne des valeurs pour un temps donné. L'idée était de comparer l'ensemble des valeurs d'un gène donné à la valeur neutre qu'aurait un échantillon de contrôle, c'est-à-dire la valeur 0 si on considère les log-ratios.

Nous avons choisi d'utiliser la construction de modèles linéaires implémentée dans le package *Limma* du logiciel Bioconductor [Gentleman *et al.*, 2004], [Reimers et Carey, 2006] car cette méthode possède les deux principaux avantages de pouvoir prendre en compte un faible nombre de réplicats et d'être applicable rapidement sur des données réelles. En effet, nous ne disposions que de deux réplicats par point de mesure dans la plupart des cas. Pour identifier les gènes dont les niveaux d'expression étaient globalement régulés, nous avons donc utilisé une analyse bayésienne empirique pour estimer les paramètres du modèle (voir page 34). Ensuite nous avons effectué un test statistique (t-test) pour évaluer dans quelle mesure chaque gène suivait le modèle. Enfin, nous avons corrigé les p-values obtenues par une correction pour les tests multiples qui permettait de contrôler le taux de faux positifs (FDR) [Hochberg et Benjamini, 1990]. Pour limiter le taux de faux positifs à 5%, nous avons pu fixer un seuil à 10^{-3} sur la p-value en suivant la méthode de Benjamini-Hochberg (voir page 34).

En utilisant cette méthode, nous avons identifié un grand nombre de gènes régulés au cours de la cinétique Cd (voir la Table 1.3). En effet, 1 330 gènes ont été identifiés comme régulés, soit 37% du génome de *Synechocystis*. Pour les autres stress étudiés, moins de 60 gènes ont été identifiés.

1.2.4 Identification des gènes répondant en deux phases

Afin de vérifier la pertinence des phases identifiées dans les réponses transcriptionnelles induites par le Cd et par H_2O_2, nous avons également choisi d'utiliser un modèle linéaire. Pour chacune des deux cinétiques, nous avons extrait deux groupes de mesures représentant chacun une des deux phases précédemment identifiées (voir Section 1.2.2). Ainsi, dans le cas du Cd, le premier groupe comprend huit réplicats (15 à 60 minutes) et le second groupe comprend 10 réplicats (90 à 360 minutes). Dans le cas de H_2O_2, le premier groupe comprend quatre réplicats (15 à 30 minutes) et le second groupe comprend six réplicats (180 à 420 minutes).

Comme précédemment, nous avons utilisé le package *Limma* du logiciel Bioconductor [Gentleman *et al.*, 2004], [Reimers et Carey, 2006] afin d'identifier les gènes dont les niveaux d'expression étaient significativement différents entre les deux phases des cinétiques. La procédure de Benjamini-Hochberg a également été appliquée afin de corriger les tests multiples avec le même seuil à 10^{-3}.

En utilisant cette méthode, nous avons identifié 791 gènes qui répondaient bien selon les deux phases décrites en réponse au stress Cd, ainsi que 228 gènes dans le cas

Stress	Cd									
Temps (min)	15	30	60	75	90	180	300	300'	360	960
Nb réplicats	2	2	4	2	2	2	2	2	2	2
Gènes induits	22	88	46	52	293	299	310	451	283	51
Gènes réprimés	8	17	10	13	250	315	328	439	310	26

TAB. 1.1 – **Influence du stress Cd sur le profil transcriptionnel.** Les cellules ont été mises en présence de Cd le temps indiqué en minutes. La mesure à 300 minutes a été répliquée biologiquement (colonnes 300 et 300'). Les gènes sont ici considérés induits si leur ratio est supérieur à 1,9 et réprimés si leur ratio est inférieur à 0,52.

Stress	H_2O_2					+Zn		+Fe		−Fe	
Temps (min)	15	30	180	300	420	30	240	240	360	2-0	1-0
Nb réplicats	2	2	2	2	2	2	2	2	2	2	2
Gènes induits	447	408	170	68	75	38	245	42	34	106	154
Gènes réprimés	490	478	92	12	55	22	221	98	100	104	109

TAB. 1.2 – **Influence des stress H_2O_2, Zn, +Fe et -Fe sur le profil transcriptionnel.** Les cellules ont été mises en présence de H_2O_2, Zn, ou Fe le temps indiqué en minutes. Pour étudier la carence en Fe, les cellules ont subi une baisse brutale de la concentration en Fe dans le milieu, de 2μM à 0μM (2-0) et de 1μM à 0μM (1-0). Les gènes sont ici considérés induits si leur ratio est supérieur à 1,9 et réprimés si leur ratio est inférieur à 0,52.

	Une phase					Deux phases	
Stress	Cd	H_2O_2	+Zn	+Fe	−Fe	Cd	H_2O_2
Induits	636	14	10	46	4	378	126
Réprimés	694	0	48	8	0	413	102
Total	1 330	14	58	54	4	791	228

TAB. 1.3 – **Nombre de gènes globalement régulés.** Les gènes globalement régulés ont été identifiés à l'aide d'un modèle linéaire en utilisant le package *Limma* du logiciel Bioconductor. Une correction pour les tests multiples a été appliquée (méthode de Benjamini-Hochberg). Finalement, seuls les gènes étant associés à une p-value inférieure à 10^{-3} ont été sélectionnés. Les gènes répondant en deux phases ont été identifiés de la même manière en considérant deux groupes de mesures caractérisant chacun une des deux phases de réponse. Les gènes montrant un profil d'expression différent entre les deux phases ont été sélectionnés.

de la réponse à H_2O_2 (voir la Table 1.3). Ceci nous a permis non seulement de confirmer la pertinence des phases proposées, mais aussi d'identifier les gènes dont le profil d'expression était significativement différent entre ces deux phases.

1.2.5 Interprétation biologique des résultats

L'étude approfondie de certaines classes de gènes a pu mettre en évidence différents résultats biologiques.

Nous avons d'abord observé, en réponse au Cd, un contrôle antagoniste des gènes impliqués dans la synthèse des protéines d'un côté, ceux-ci étant réprimés, et des gènes impliqués dans la maturation et la dégradation des protéines de l'autre côté, ceux-ci étant induits. Ainsi, les gènes codant les protéines ribosomales sont largement réprimés (voir Annexe A). Rappelons que ces complexes de protéines participent à la synthèse des protéines. D'un autre côté, les gènes codant les protéines chaperones et les proteases sont fortement induits. Or, ces classes de protéines sont impliquées dans la maturation et la dégradation des protéines.

D'autre part, nous avons noté que la plupart des gènes codant les aminoacyl-tRNA synthetases n'étaient pas significativement régulés. En l'absence du stress Cd, ces gènes sont modérément exprimés, alors que les gènes codant les protéines ribosomales sont fortement exprimés. En présence du stress Cd, nous avons noté que les gènes codant les aminoacyl-tRNA synthetases ne sont pas régulés, alors que les gènes codant les protéines ribosomales sont fortement réprimés. Par conséquent, nous en avons conclu que les mécanismes de régulation de l'expression des gènes en réponse au stress Cd favorisaient la répression des gènes codant les protéines ribosomales, car ces gènes sont d'habitude fortement exprimés ; ceci représente une forte charge métabolique. Ainsi, la régulation de la transcription en réponse au stress Cd limite l'expression des gènes dont la charge métabolique est habituellement forte.

Dans le cas de la réponse à H_2O_2, nous avons également observé une répression des gènes codant les protéines ribosomales et une absence de régulation majeure des gènes codant les aminoacyl-tRNA synthetases.

Cette analyse nous a permis entre autres de définir deux phases principales lors des réponses transcriptionnelles aux stress induits par le Cd et l'H_2O_2. Nous avons identifié les gènes dont le profil d'expression était significativement différent au cours de ces deux phases. Nous avons identifié également les gènes globalement induits ou réprimés lors des différentes réponses transcriptionnelles étudiées. Par conséquent, nous avons mis en évidence les deux principales tendances en termes de réponse transcriptionnelle, la réponse uniforme et la réponse en deux phases. De plus, nous proposons l'hypothèse suivante, selon laquelle les mécanismes de régulation de la transcription tendraient à limiter la charge métabolique représentée par l'expression de certains gènes en les réprimant.
À ce stade, nous avons alors voulu regrouper les gènes co-exprimés pour chacune de ces tendances. Pour cela, nous avons étudié la similarité entre les profils d'expression des gènes.

1.3 Identification des classes de gènes co-exprimés

Pour regrouper les gènes co-exprimés, nous avons choisi d'étudier les profils d'expression des gènes, et en particulier leur similarité. L'une des approches les plus couramment utilisées pour identifier des classes de gènes ayant des profils d'expression similaires est la classification hiérarchique [Eisen *et al.*, 1998]. Cette méthode permet d'obtenir plusieurs partitions possibles de l'ensemble des gènes. Cela implique entre autres que chaque gène appartient à une et une seule classe. Pourtant, un gène peut avoir plusieurs fonctions et par conséquent faire partie de plusieurs classes.

C'est pourquoi une autre approche a été développée en utilisant une classification pyramidale (voir page 38). Cette approche est une généralisation des hiérarchies (voir page 37). Elle permet notamment d'obtenir des classes éventuellement recouvrantes, et donc de tenir compte du fait qu'un gène peut avoir plusieurs fonctions. Toutefois, la complexité de l'algorithme qui réalise cette classification est telle qu'il ne peut être utilisé que sur un ensemble de départ d'au plus 250 individus.

Un modèle composite de classification a été proposé notamment pour palier ce problème de complexité (voir page 40). Ce modèle composite permet de combiner deux niveaux de flexibilité différents en se basant sur les hiérarchies et les pyramides. En effet, le modèle pyramidal est plus contraint et son utilisation locale permet de préciser l'ordre entre les objets.

À ce stade, nous avons alors voulu développer une méthode de classification mixte hiérarchique-pyramidale en nous basant sur ce modèle composite. Pour cela, nous avons d'abord adapté le modèle composite afin de l'appliquer au problème de la classification des profils de gènes. Ainsi, une hiérarchie est d'abord construite en partant de l'ensemble des gènes à classer, puis elle est découpée afin de définir des classes. Ces classes sont ensuite analysées plus finement à l'aide de pyramides. Pour réaliser cela, nous avons automatisé le découpage de la hiérarchie afin de développer une méthode de classification mixte hiérarchique-pyramidale que nous avons ensuite appliquée aux données d'expression chez *Synechocystis*.

1.3.1 Adaptation du modèle composite de classification

L'objectif était d'adapter le modèle composite afin de développer une méthode de classification mixte hiérarchique-pyramidale permettant d'identifier des classes de gènes éventuellement recouvrantes et applicable aux données biologiques. L'ensemble des individus est constitué des gènes que nous voulons classer. La méthode s'articule en trois étapes.

Dans un premier temps, nous avons défini une matrice de dissimilarité entre les gènes. Puis, nous avons construit une hiérarchie à partir de ces gènes et de leur dissimilarité.

Dans un deuxième temps, la hiérarchie a été découpée à un niveau d'indice donné de manière à obtenir une partition de l'ensemble de départ (voir Figure 1.1). En effet, si nous fixons une valeur seuil d'indice h de la hiérarchie, nous pouvons alors considérer la section de la hiérarchie à cette hauteur h. Toutes les classes dont l'indice est supérieur à h sont supprimées. Il nous reste alors un certain nombre n de nouvelles racines qui définissent

des arbres disjoints. Chacun des ces n arbres engendre un sous-ensemble d'objets de l'ensemble de départ, autrement dit une classe. Ainsi, une partition de l'ensemble des objets à étudier est obtenue.

Dans un troisième temps, une pyramide a été construite pour chacune des n classes de gènes identifiées pour permettre une analyse plus fine des relations entre les gènes (voir Figure 1.1).

1.3.2 Automatisation du découpage de la hiérarchie

Une des questions principales de la méthode de classification mixte hiérarchique-pyramidale est de savoir comment déterminer le seuil de découpage de la hiérarchie. Notre but était d'automatiser ce choix. Pour cela, nous avons défini des critères tenant compte de la qualité globale de la partition obtenue pour un seuil de découpage donné, ainsi qu'un critère technique concernant l'algorithme de construction des pyramides.

1.3.2.1 Définition de critères de partitionnement

La question principale était donc de déterminer la valeur de l'indice à fixer pour le découpage de la hiérarchie et l'obtention de la partition. Il est clair que plus cette valeur est élevée et plus le nombre de classes est faible. Mais la taille de ces classes en est d'autant plus grande. Or, nous sommes confrontés à une limitation technique puisque l'algorithme de construction des pyramides, la CAP (voir page 38), ne peut s'appliquer qu'à 250 individus au plus.

La valeur maximale est celle de l'indice de la racine. Dans ce cas, la partition possède une classe unique qui est l'ensemble de départ. Par conséquent, ce découpage ne présente aucun intérêt. A l'inverse, une valeur d'indice proche de 0 se situe près des feuilles dont l'indice est nul par convention. Le nombre de classes est alors très élevé, proche du nombre total d'individus, et la taille des classes est très faible.

À ce stade, l'objectif était donc de définir un critère de partitionnement pour indiquer la valeur d'indice permettant d'obtenir la meilleure partition. Pour cela, nous avons choisi de quantifier la qualité d'une partition par la *distance intra-classe* d'une part et la *distance inter-classes* de l'autre. La distance intra-classe caractérise l'homogénéité de chaque classe. Par conséquent, la distance intra-classe globale d'une partition est d'autant plus faible que les éléments de chaque classe sont proches les uns des autres. La distance inter-classes, quant à elle, caractérise l'éloignement des classes les unes par rapport aux autres. Ainsi, une partition dont les classes sont très espacées a une distance inter-classes élevée. Notre objectif était d'obtenir une partition telle que ses classes étaient homogènes et éloignées les unes des autres. Par conséquent, nous avons choisi de maximiser la distance inter-classes et de minimiser la distance intra-classe. Nous avons utilisé les critères de Dunn et Davies-Bouldin définis par Bolshakova *et al.* comme suit [Bolshakova et Azuaje, 2003] :

$$D(U) = \min_{1\leq i\leq c}\left\{\min_{1\leq j\leq c, j\neq i}\left\{\frac{\delta(X_i, X_j)}{\max_{1\leq k\leq c}\{\Delta(X_k)\}}\right\}\right\} \tag{1.3}$$

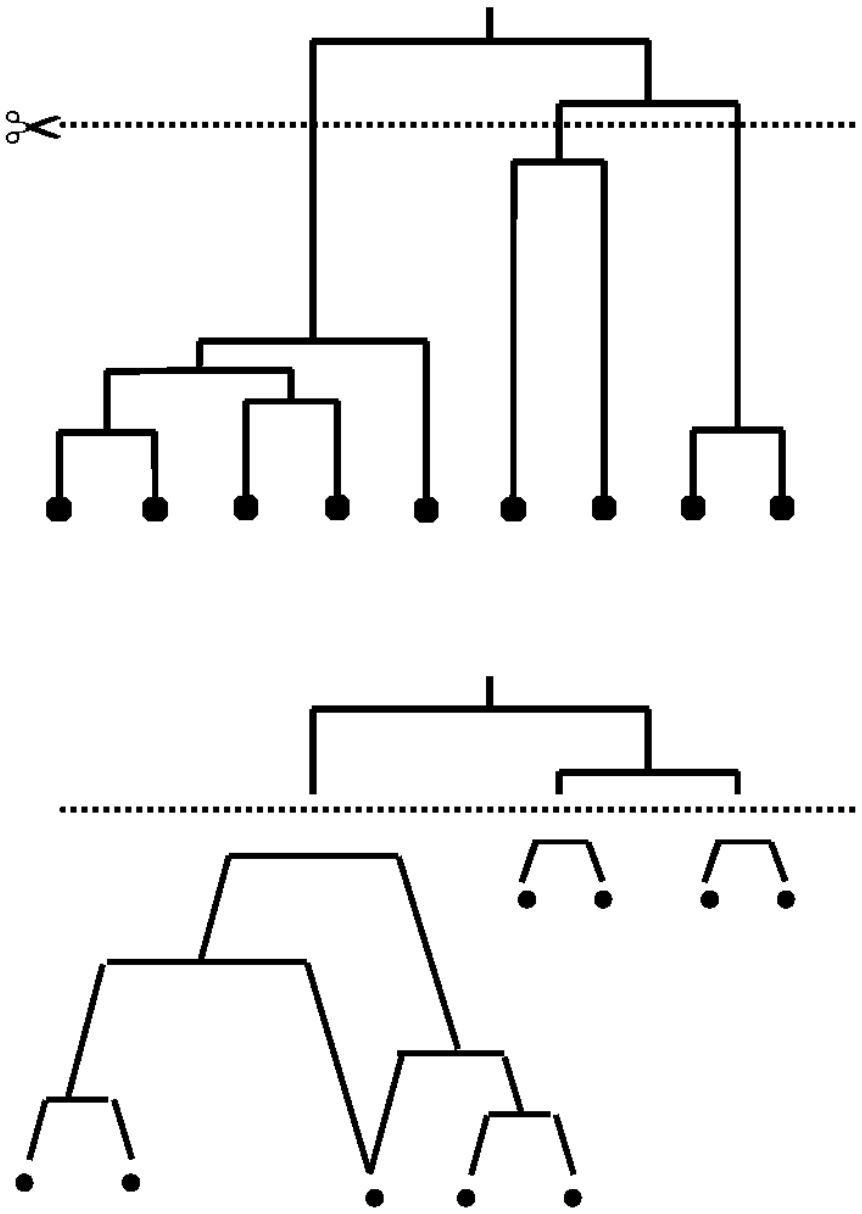


FIG. 1.1 – **Principe de la classification mixte hiérarchique-pyramidale.** Cette figure illustre le principe de la classification mixte hiérarchique-pyramidale. Une hiérarchie est d'abord construite. Elle est ensuite découpée à une hauteur donnée de manière à obtenir une partition de l'ensemble de départ. Enfin, une pyramide est construite pour chaque classe de la partition.

où $\delta(X_i, X_j)$ définit la distance inter-classes entre X_i et X_j ; $\Delta(X_k)$ représente la distance intra-classe de la classe X_k, et c est le nombre de classes de la partition U. Nous avons donc cherché à maximiser cet indice.

$$DB(U) = \frac{1}{c} \sum_{i=1}^{c} \max_{i \neq j} \left\{ \frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)} \right\} \tag{1.4}$$

où $\delta(X_i, X_j)$ définit la distance inter-classes entre X_i et X_j ; $\Delta(X_k)$ représente la distance intra-classe de la classe X_k, et c est le nombre de classes de la partition U. Nous avons donc cherché à minimiser cet indice.

L'indice *Silhouette* n'a pas été utilisé car les auteurs ont montré que sa validité était discutable en termes de biologie [Bolshakova et Azuaje, 2003]. De plus, les auteurs ont montré que l'utilisation de différentes distances (Δ et δ) n'avait pas d'influence majeure sur les critères. Ainsi, nous avons choisi d'implémenter ces critères en utilisant la distance euclidienne.

1.3.2.2 Développement d'une méthode de découpage automatique d'une hiérarchie

Le but de cette méthode était d'utiliser une hiérarchie pour partitionner l'ensemble des individus considérés, afin de créer des classes de taille compatible avec l'algorithme de CAP et de meilleure qualité possible, en se basant sur les critères de taille et de partitionnement. Pour cela, nous avons suivi un parcours de la hiérarchie par indice décroissant, c'est-à-dire en partant de la racine et en descendant vers les feuilles. Un nombre de classes à étudier était fixé au départ comme paramètre. En effet, le but n'était pas d'obtenir directement des classes de petite taille mais plutôt de sélectionner des groupes d'individus assez différenciés, homogènes, quitte à recouper ceux qui contenaient trop d'individus. Néanmoins, il pourrait être intéressant de tester d'autres approches plus fines. Une idée serait par exemple de calculer une valeur d'indice pour certaines partitions régulièrement espacées, et d'en déduire, par une méthode de gradient, la zone de l'optimum du critère considéré.

Pour une hauteur donnée, nous avons donc testé la coupe de la hiérarchie à cette hauteur. Pour cela, nous avons considéré la qualité de la partition obtenue en calculant les indices de Dunn et Davies-Bouldin. Les classes qui optimisaient les différents critères étaient alors retenues.

À ce stade, si les critères sélectionnaient la même hauteur, cette dernière était choisie pour le découpage. Sinon, parmi les hauteurs qui optimisaient les critères, nous avons choisi de conserver la plus élevée dans la hiérarchie car ceci permettait de découper le moins possible la hiérarchie de départ, tout en sachant que les classes trop grandes étaient de toute façon redécoupées par la suite jusqu'à avoir un nombre d'individus inférieur au seuil fixé.

Cette méthode permet donc de découper automatiquement une hiérarchie en respectant strictement le critère de taille et en optimisant le critère de partitionnement.

1.3.3 Application de la classification mixte

Nous avons appliqué cette méthode de classification aux données transcriptome obtenues en réponse aux différents stress étudiés chez *Synechocystis* afin de classer les gènes. Nous considérons pour cela la distance euclidienne entre les valeurs d'expression des gènes. Dans un premier temps, nous avons étudié les gènes globalement régulés. Dans un second temps, nous avons étudié les réponses transcriptionnelles en deux phases mises en évidence dans le cas du Cd et de l'H_2O_2.

1.3.3.1 Classification des gènes globalement régulés

Dans le cas de la réponse au stress induit par le Cd, nous avons identifié 1 330 gènes globalement régulés (voir Section 1.2.3). Nous avons alors retiré les gènes qui avaient des valeurs manquantes provenant de problèmes techniques (hybridation des puces ; spots de mauvaise qualité) n'ayant pas permis d'obtenir des valeurs pour le niveau d'expression. Nous avions alors un ensemble de 1 264 gènes, représenté par une matrice des modalités de $1\,264 \times 22$ (voir page 36). En effet, nous avions pour ce jeux de données 22 réplicats décrivant les deux phases. Puis, nous avons calculé la matrice de distance entre ces gènes.

Nous avons alors appliqué la méthode de classification mixte hiérarchique pyramidale à ces données. Nous avons obtenu une hiérarchie qui a ensuite été découpée pour obtenir huit classes de gènes pour lesquelles nous avons construit des pyramides. À partir des 1 264 gènes de départ, nous avons ainsi construit huit pyramides contenant 62, 238, 97, 211, 218, 140, 148 et 150 gènes (voir la Figure 1.2 et l'Annexe D).

Nous avons observé que les deux premières classes reflétaient les deux tendances majeures de comportement, à savoir l'induction (profil positif) et la répression (profil négatif). L'ensemble des 608 gènes induits étaient séparés en quatre classes (les classes 1, 2, 7 et 8 sur la Figure 1.2). Les gènes réprimés étaient également séparés en quatre classes (les classes 3, 4, 5 et 6 sur la Figure 1.2). Certaines classes montraient des gènes dont les profils étaient proches, comme les classes 7 et 8. La classe d'origine était effectivement relativement homogène mais a du être découpée car sa taille (308 gènes) excédait le seuil de 250 autorisé pour réaliser une pyramide.

1.3.3.2 Classification des gènes répondant en deux phases

Deux phases de réponse ont pu être identifiées pour les deux cinétiques Cd et H_2O_2. Nous avons donc appliqué la classification mixte à ces deux jeux de données.

Dans le cas de la réponse au stress induit par le Cd, nous avons identifié 791 gènes répondant en deux phases (voir Section 1.2.3). Puis, nous avons retiré les gènes qui avaient des valeurs manquantes comme précédemment (voir Section 1.3.3.1). Nous avions alors une matrice des modalités de 754×18 (voir page 36) que nous avons utilisée pour construire la matrice de distance.

L'application de la méthode de classification mixte hiérarchique pyramidale à ces données nous a permis d'obtenir une hiérarchie découpée en cinq classes. À partir des 754 gènes de départ, nous avons ainsi construit cinq pyramides contenant 43, 172, 161, 228 et 150 gènes (voir la Figure 1.3).

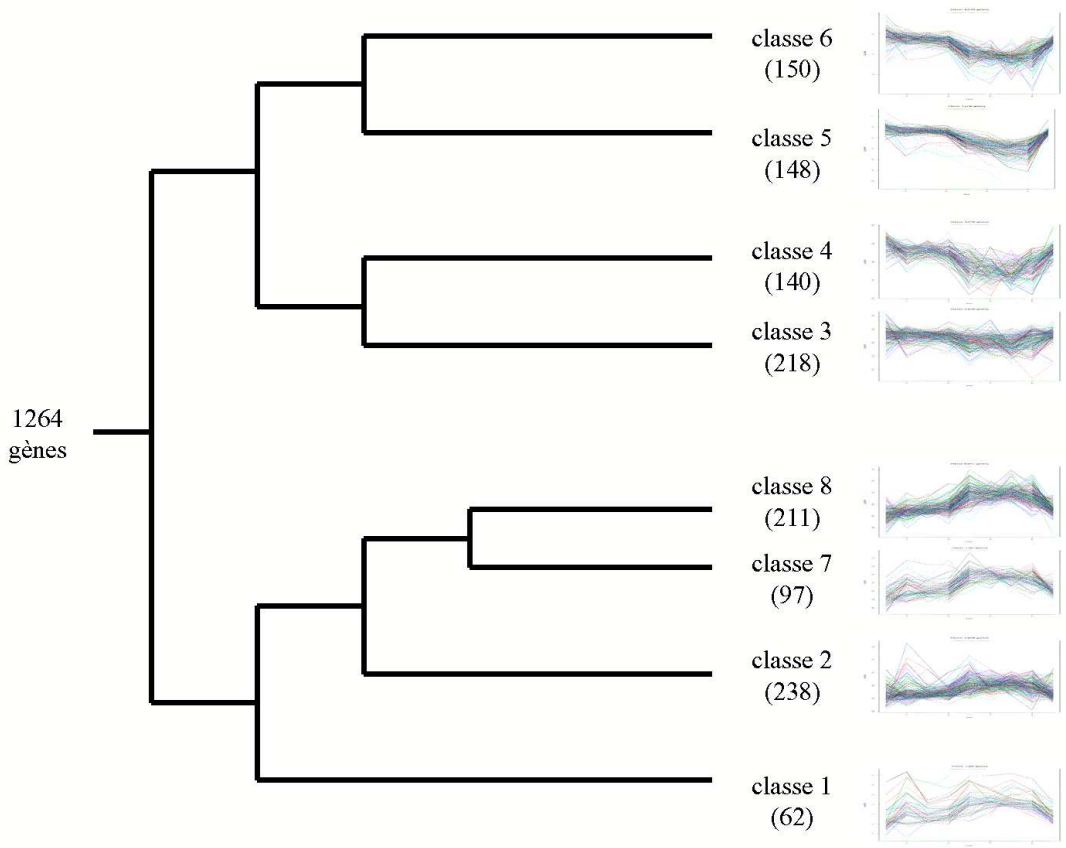


FIG. 1.2 – **Classes de gènes globalement régulés pour le Cd.** Cette figure représente les huit classes identifiées dans l'ensemble des gènes globalement régulés au cours de la réponse au stress Cd. Dans la partie droite, chaque encadré représente les profils d'expression des gènes appartenant à cette classe.

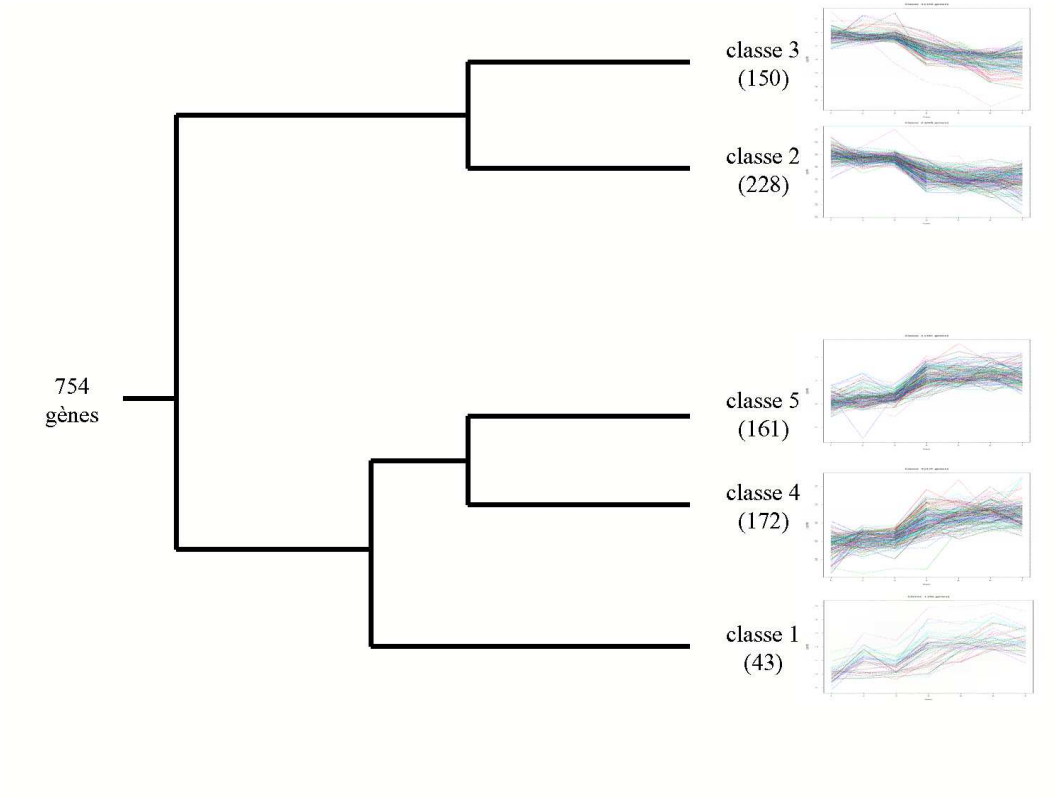


FIG. 1.3 – **Classes de gènes répondant en deux phases pour le Cd.** Cette figure représente les différentes classes identifiées dans l'ensemble des gènes répondant en deux phases au cours de la réponse au stress Cd. Dans la partie droite, chaque encadré représente les profils d'expression des gènes appartenant à cette classe.

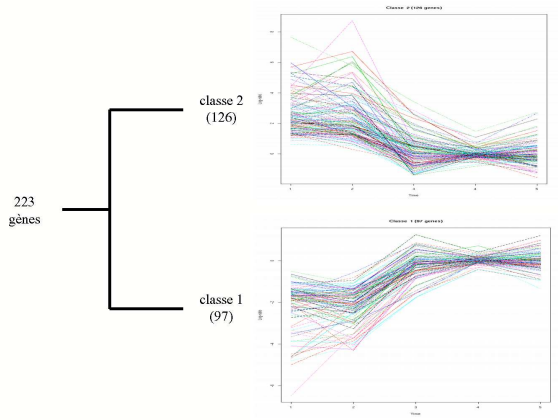


FIG. 1.4 – **Classes de gènes répondant en deux phases pour le H_2O_2.** Cette figure représente les différentes classes identifiées dans l'ensemble des gènes répondant en deux phases au cours de la réponse au stress H_2O_2. Dans la partie droite, chaque encadré représente les profils d'expression des gènes appartenant à cette classe.

De même que précédemment, nous avons observé que les deux premières classes reflétaient les deux tendances majeures de comportement, à savoir l'induction lors de la phase massive (profil nul puis positif) et la répression au cours de cette même phase (profil nul puis négatif). L'ensemble des 376 gènes induits étaient séparés en trois classes (les classes 1, 4, et 5 sur la Figure 1.3). Les gènes réprimés étaient séparés en deux classes (les classes 2 et 3 sur la Figure 1.3). Certaines classes montraient des gènes dont les profils étaient proches comme les classes 4 et 5. La classe d'origine était effectivement relativement homogène mais a du être découpée pour les mêmes raisons que précédemment (333 gènes).

Dans le cas de la réponse au stress induit par le H_2O_2, nous avons identifié 228 gènes répondant en deux phases (voir Section 1.2.3). L'application de la méthode de classification mixte hiérarchique pyramidale à la matrice des modalités de taille 223×10 a produit deux pyramides contenant 97 et 126 gènes (voir la Figure 1.4). Ces deux classes représentaient de manière attendue les gènes induits au cours de la phase massive (classe 2) et les gènes réprimés au cours de la phase massive (classe 1).

1.3.3.3 Exemple de classification

Nous présentons en exemple les classifications obtenues pour les 62 gènes de la classe 1 globalement induits au cours de la cinétique Cd (voir Figure 1.2). Ces figures représentent une hiérarchie (voir Figure 1.5) et une pyramide (voir Figure 1.6). Comme précédemment expliqué, l'ordre induit par la pyramide sur les gènes est beaucoup plus contraint que ne l'est celui induit par la hiérarchie. En effet, les gènes de cette pyramides sont ordonnés de façon unique alors qu'un grand nombre d'ordres différents peuvent être obtenus par la hiérarchie en faisant pivoter les ensembles de gènes autour des nœuds.

La hiérarchie permet d'identifier trois principales classes de gènes (bleu, vert et orange sur la figure 1.5). A l'intérieur de ces classes, nous avons défini des groupes de deux à six gènes en utilisant le structure de la hiérarchie. À quelques exceptions près, ces groupes de gènes se retrouvent dans la pyramide (voir Figure 1.6) mais l'ordre des groupes a été changé et fixé par la pyramide.

À titre d'illustration, nous avons considéré les gènes codant des protéines ribosomales. Les groupes de 18 gènes mis en évidence en bas de la hiérarchie et de la pyramide contiennent les même gènes. Notons cependant que les contraintes de la pyramide ont permis de rapprocher les deux gènes *sll0898* et *slr0551* du gène *slr0853* codant également une protéine ribosomale. En revanche, le lien entre ces gènes n'est pas évident à partir de la hiérarchie. Le gène *slr0853* est beaucoup plus éloigné mais il pourrait être localisé à différents endroits selon la représentation choisie. En particulier, la situation représentée par la pyramide pourrait être obtenue avec différentes permutations, ramenant le gène *slr0853* vers le bas et les deux gènes *sll0898* et *slr0551* en haut du groupe. Ceci illustre l'intérêt des pyramides où les contraintes sur les paliers permettent de choisir un ordre parmi différents possibles pour la hiérarchie.

Notons de manière plus générale que la méthode de classification mixte peut être appliquée sur une distance quelconque entre les gènes. Ainsi, il pourrait être intéressant

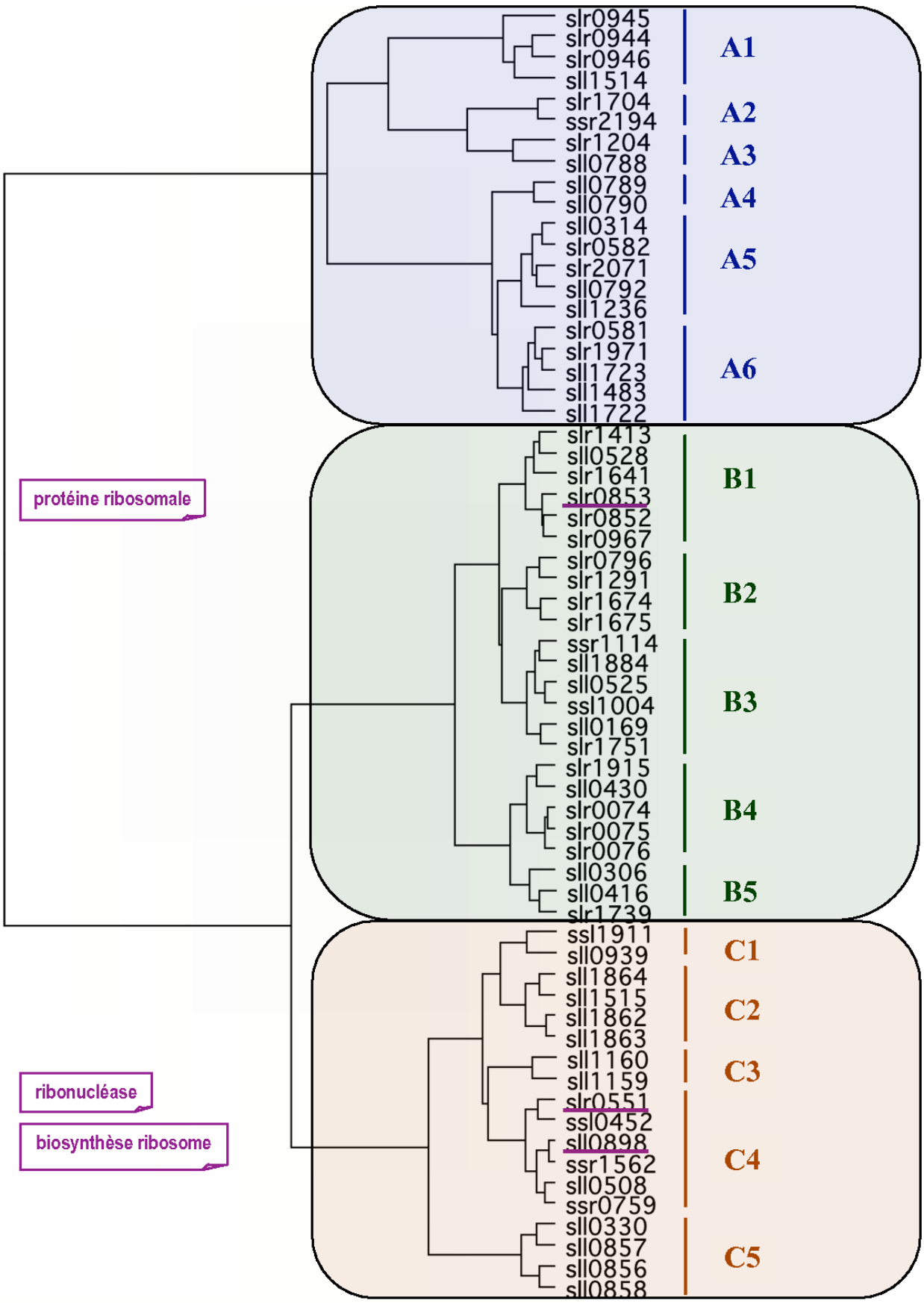


FIG. 1.5 – **Hiérarchie de la classe des gènes induits.** Cette figure représente la hiérarchie obtenue sur les 62 gènes de la classe 1 induits en réponse au Cd.

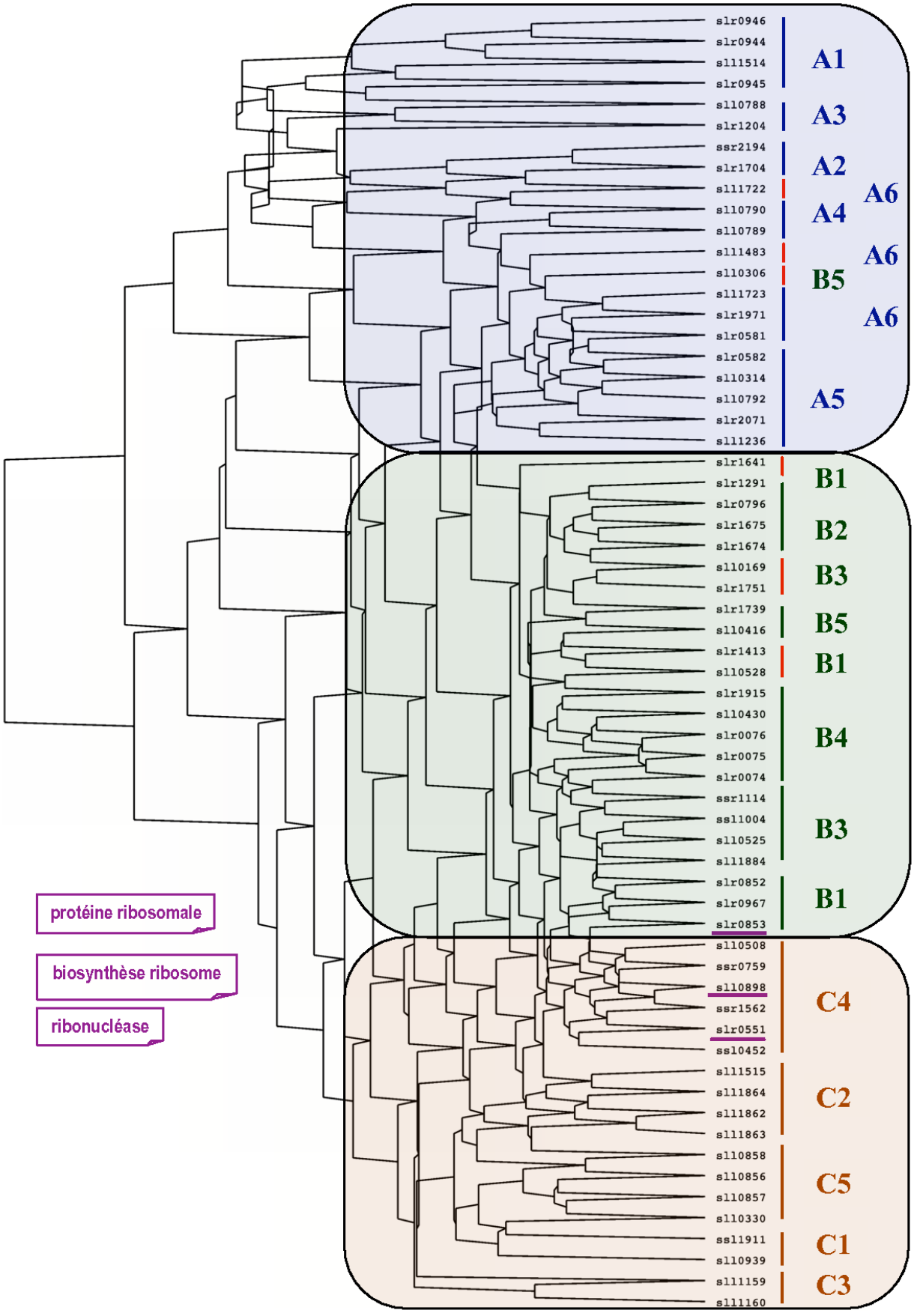


FIG. 1.6 – **Pyramide de la classe des gènes induits.** Cette figure représente la pyramide obtenue sur les 62 gènes de la classe 1 induits en réponse au Cd.

d'ajouter une information fonctionnelle ou bien une information sur la localisation des gènes dans le génome. Enfin, la problématique du découpage automatique s'inscrit dans un domaine plus large que celui présenté ici. En effet, il est intéressant pour n'importe quelle hiérarchie de définir des critères permettant de déterminer de manière automatique le seuil le plus pertinent pour découper la hiérarchie et ainsi obtenir une partition optimale des éléments. Ceci est indépendant de ce qu'on veut faire après avec les classes.

Nous avons développé une méthode de classification mixte hiérarchique-pyramidale basée sur le découpage automatique d'une hiérarchie. De plus, nous avons appliqué cette méthode à la classification des gènes de *Synechocystis* en nous basant sur la similarité des profils d'expression de ces gènes en réponse à différents stress.
À ce stade, nous avons voulu utiliser ces classes de gènes pour tenter d'expliquer pourquoi certaines classes étaient induites et d'autres réprimées. Pour cela, nous avons voulu mettre en évidence d'éventuels biais de composition entre différentes classes de gènes ou de protéines.

1.4 Mise en évidence de biais de composition

Comme expliqué dans l'étude bibliographique (voir page 42), Fauchon *et al.* ont mis en évidence, chez la levure, un mécanisme d'économie des acides aminés soufrés, en réponse à un stress induit par le Cd [Fauchon *et al.*, 2002]. Ainsi, les gènes codant des protéines dont la composition en acides aminés soufrés était importante avaient tendance à être réprimés, alors qu'au contraire, les gènes codant des protéines à faible teneur en acides aminés soufrés avaient tendance à être induits. Le soufre est en effet redirigé vers la voie de biosynthèse du glutathion qui est utile à la détoxication du cadmium. À ce stade, nous voulions savoir si ce phénomène était observable également chez *Synechocystis*.

De plus, nous nous sommes posé plus généralement la question de la relation entre les propriétés des gènes, ou des protéines qu'ils codent, et leur niveau d'expression. Pour étudier cela, nous avons développé une méthode générale de détection automatique de biais entre deux groupes de gènes ou de protéines. Nous avons d'abord appliqué cette méthode à la levure dans le but de retrouver l'hypothèse de départ concernant le biais en soufre. Ensuite, nous avons appliqué cette méthode à *Synechocystis*. Enfin, nous avons implémenté cette méthode sous la forme d'un outil appelé BiasSeeker.

1.4.1 Développement d'une méthode de détection automatique de biais de composition

Après avoir développé une méthode de détection de biais, nous l'avons appliquée à la levure afin de retrouver le biais en soufre précédemment identifié.

1.4.1.1 Présentation de la méthode

L'idée de départ était d'étudier l'existence d'un mécanisme d'économie du soufre chez *Synechocystis*. Pour cela, nous voulions comparer la composition des protéines en acides aminés soufrés entre deux populations caractéristiques de la régulation de la transcription lors d'un stress induit par le Cd. Le premier groupe comprenait l'ensemble des protéines codées par les gènes induits. Le second groupe était constitué, pour sa part, des protéines codées par les gènes réprimés.

Nous avons étendu cette question en considérant différentes propriétés de composition des gènes et des protéines (composition en acides aminés ou en atomes spécifiques), ainsi que des propriétés biochimiques telles que l'hydrophobicité ou le point iso-électrique (voir section 1.4.1.2). Par ailleurs, nous avons généralisé cette approche en considérant deux populations quelconques.

La méthode que nous avons développée se décompose en trois étapes principales : le calcul des valeurs du paramètre étudié, la comparaison de ces valeurs entre les deux groupes considérés et la détermination de l'existence d'un biais. Prenons pour exemple le biais en soufre.

Dans un premier temps, nous avons calculé la valeur du paramètre pour chaque protéine. Dans le cas du biais en soufre, la valeur était le nombre d'atomes de soufre de la protéine.

Ensuite, nous avons comparé les deux groupes de protéines. Pour cela, nous avons choisi d'effectuer un test statistique. Nous avions alors à disposition deux groupes de réalisations d'une variable aléatoire qui suivait la même loi de distribution pour les deux groupes par hypothèse. Nous avons choisi comme hypothèse nulle l'identité des moyennes. Nous avons alors décidé d'effectuer un test de Wilcoxon car il présentait l'avantage d'être non paramétrique, c'est-à-dire de ne supposer aucune loi de distribution pour la variable aléatoire considérée. Le principe de ce test est rappelé en annexe C.1.

Enfin, nous avons pu fixer un seuil d'erreur (par exemple $\alpha = 10^{-3}$) de manière à déterminer si l'hypothèse pouvait être rejetée ou non. Dans le cas d'un rejet, l'identité des moyennes est mise en défaut et nous en concluons qu'un biais existe pour ce paramètre en question entre les deux groupes de protéines.

À cette étape, le seuil d'erreur n'est pertinent que si nous effectuons une correction pour les tests multiples puisque nous réalisons plusieurs tests d'hypothèse. Néanmoins, les méthodes présentées à ce propos telles que le contrôle du taux de faux positifs (FDR) supposent l'indépendance des tests effectués (voir page 34). Or, dans ce cas précis, les biais sont liés les uns aux autres et les tests sont donc dépendants. En effet, les compositions en acides aminés et en atomes sont dépendantes. De même, la charge de la protéine est déterminée par la composition en acides aminés. Ainsi, la probabilité d'observer un biais en termes de charge est différente si un biais en glutamate (acide aminé chargé positivement) est observé ou non. Par conséquent, ces méthodes de correction sont difficilement applicables [Benjamini et Yekutieli, 2001].

1.4.1.2 Présentation des paramètres étudiés

Nous avons classé les différents paramètres étudiés en quatre classes selon qu'ils caractérisaient les propriétés générales, les propriétés des acides aminés, la composition en acides aminés ou encore la composition en atomes des protéines. La liste des acides aminés est rappelée en annexe B.1 et l'ensemble des paramètres étudiés est présenté en détail dans l'annexe B.2.

Les sept propriétés générales étudiées sont les suivantes : la longueur ; le poids moléculaire ; le point isoélectrique, c'est-à-dire la valeur de pH pour laquelle la protéine est neutre ; le caractère aliphatique, qui intervient dans la thermostabilité de la protéine ; les propriétés optiques, en particulier l'absorbtion de la lumière ; le caractère hydrophobe.

Nous avons également considéré les propriétés des acides aminés qui sont représentées sur la figure 1.7. De plus, nous avons étudié les compositions en acides aminés (20 acides aminés naturels) et en atomes (carbone, azote, oxygène, soufre, hydrogène).

1.4.1.3 La cinétique Cd chez *S. cerevisiae*

Comme expliqué précédemment (voir Section 1.4), la problématique avait été introduite suite à la mise en évidence, chez la levure, d'un mécanisme d'économie des acides aminés soufrés en réponse à un stress au cadmium. Par conséquent, nous avons voulu vérifier si notre méthode permettait bien d'identifier ce biais en soufre. Pour cela, nous avons récupéré la liste des gènes qui avaient été sélectionnés comme induits ou réprimés dans l'article de référence. Après avoir collecté les séquences des protéines correspondantes, nous avons appliqué notre méthode d'identification de biais.

Les résultats de cette analyse étaient parfaitement conformes à ce que nous attendions. En effet, un biais dans la composition en soufre a été mis en évidence (voir Table C.1 de l'Annexe C). Ainsi, les protéines induites contenaient moins de soufre que les protéines réprimées (voir Figure 1.8). De plus, un biais dans la composition en acides aminés a permis de montrer que cette différence provenait essentiellement des cystéines. En effet, les deux seuls acides aminés contenant du soufre sont la cystéine et la méthionine. Aucun biais dans la composition en méthionine n'étant apparu lors de ces analyses, nous en avons conclu que le biais en soufre provenait essentiellement d'un biais dans la composition en cystéines. Ainsi, les protéines induites contenaient moins de cystéines que les protéines réprimées (voir Figure 1.9).

Ceci nous a permis de valider la méthode dans une certaine mesure. Notons toutefois que cette méthode ne tient pas compte de la relation entre les ARNm et les protéines. Nous supposons que la quantité de protéines est corrélée à la quantité d'ARNm. De plus, nous ne tenons pas compte du taux d'induction des différents ARNm puisque nous avons simplement considéré deux groupes selon qu'ils étaient induits ou réprimés. Cependant, cette méthode peut permettre d'identifier des biais et aider à générer de nouvelles hypothèses qui devront être validées expérimentalement par la suite. Nous avons alors voulu l'appliquer à *Synechocystis*.

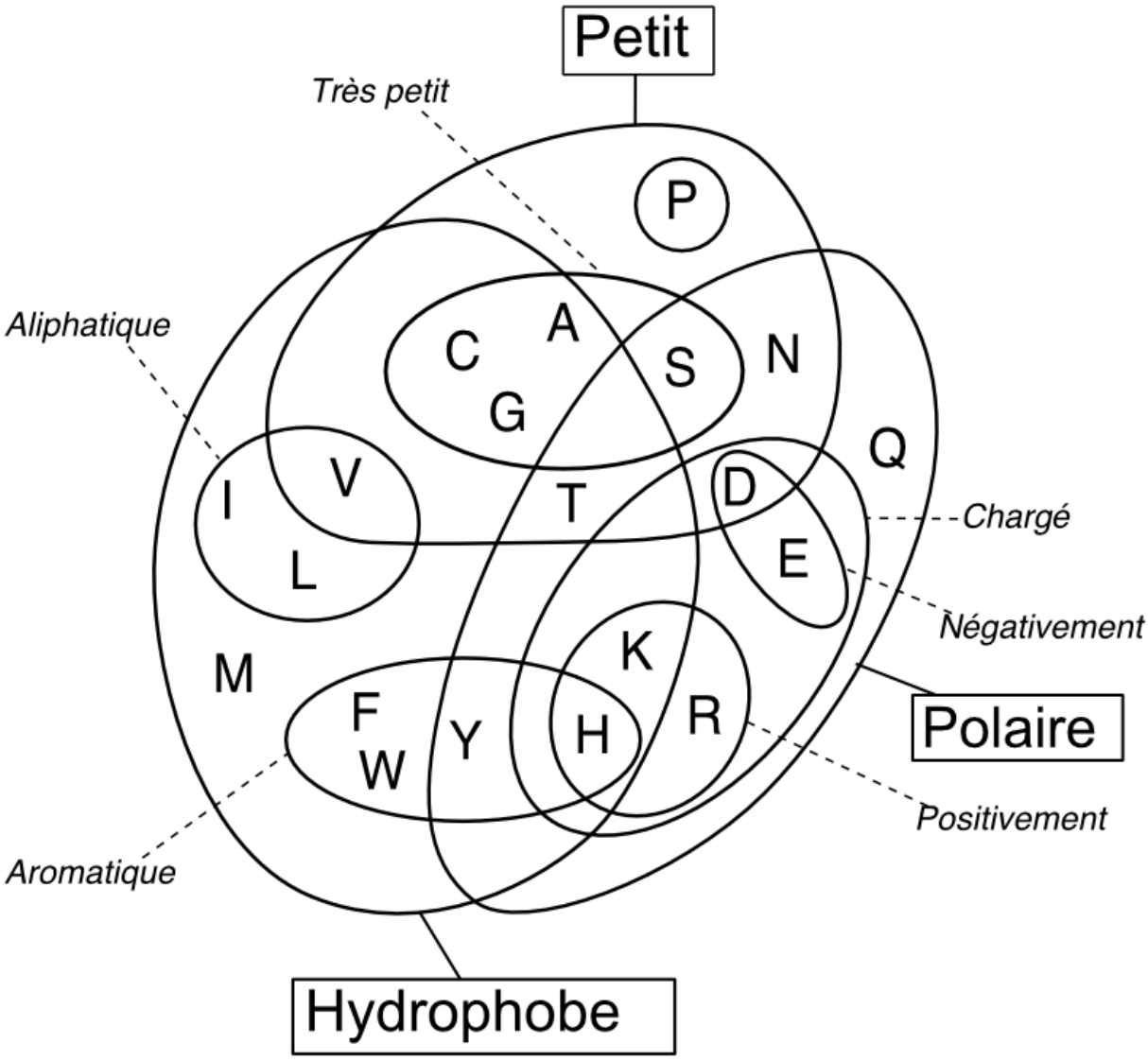


FIG. 1.7 – **Classification des acides aminés.** Ce diagramme met en évidence les différentes caractéristiques des 20 acides aminés naturels. La figure est une adaptation de travaux antérieurs [Livingstone et Barton, 1993], [Launay, 2007].

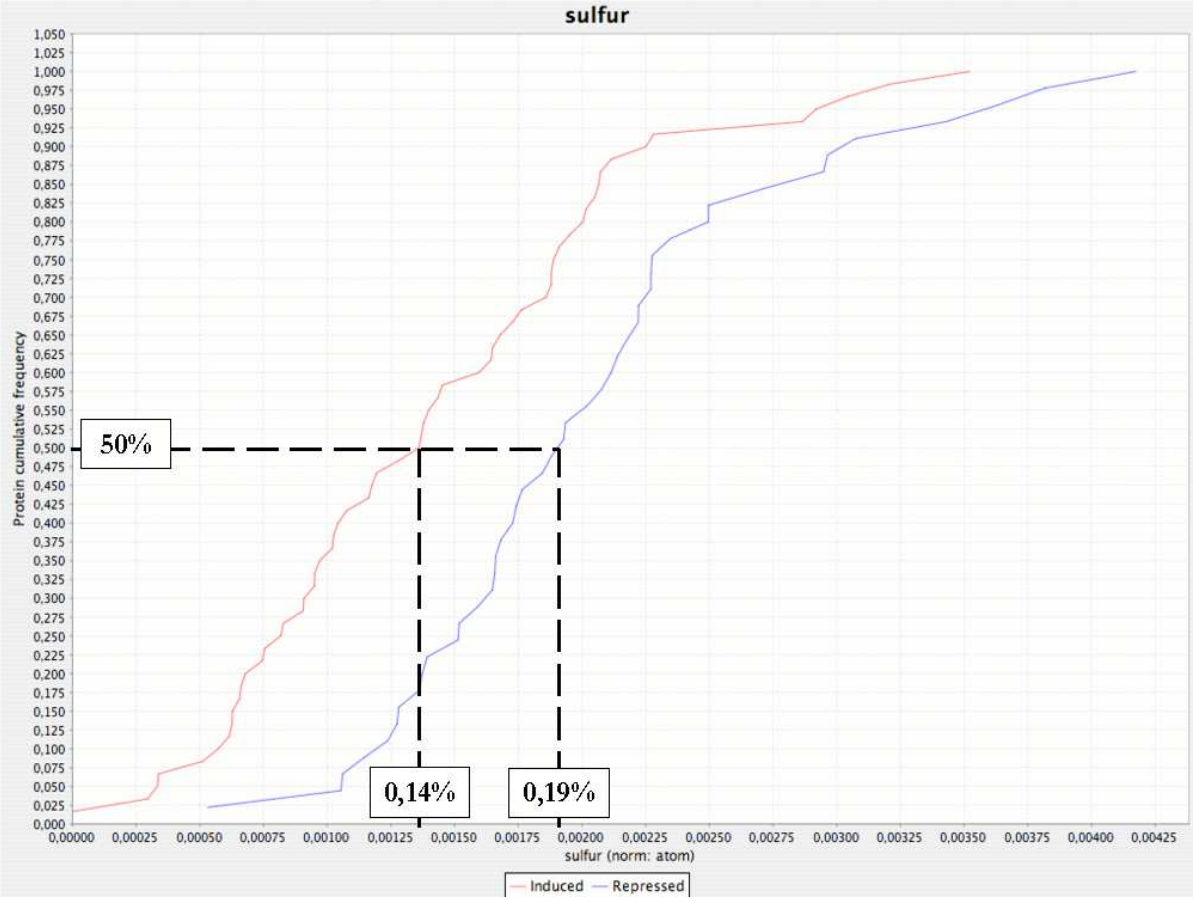


FIG. 1.8 – **Biais en soufre chez la levure.** Ce graphique représente les fréquences cumulées des protéines possédant un pourcentage donné d'atomes de soufre [Baudouin-Cornu *et al.*, 2001]. Les abscisses représentent en effet le pourcentage d'atomes de soufre de la protéine (le nombre d'atome de soufre a été normalisé par le nombre d'atomes total de la protéine). Les deux courbes montrent les fréquences cumulées pour les deux groupes considérés, c'est-à-dire les 60 protéines correspondant aux gènes induits et les 45 protéines correspondant aux gènes réprimés. Prenons un exemple pour illustrer la lecture de ce graphique : dans le cas des induits (en rouge), 50% des protéines contiennent moins de 0,14% de soufre ; dans le cas des réprimés (en bleu), 50% des protéines contiennent moins de 0,19% de soufre. De manière globale, ce graphique montre que les induits possèdent un plus faible pourcentage d'atomes de soufre. Notons qu'il ne permet en rien de conclure quant à l'existence d'un biais significatif.

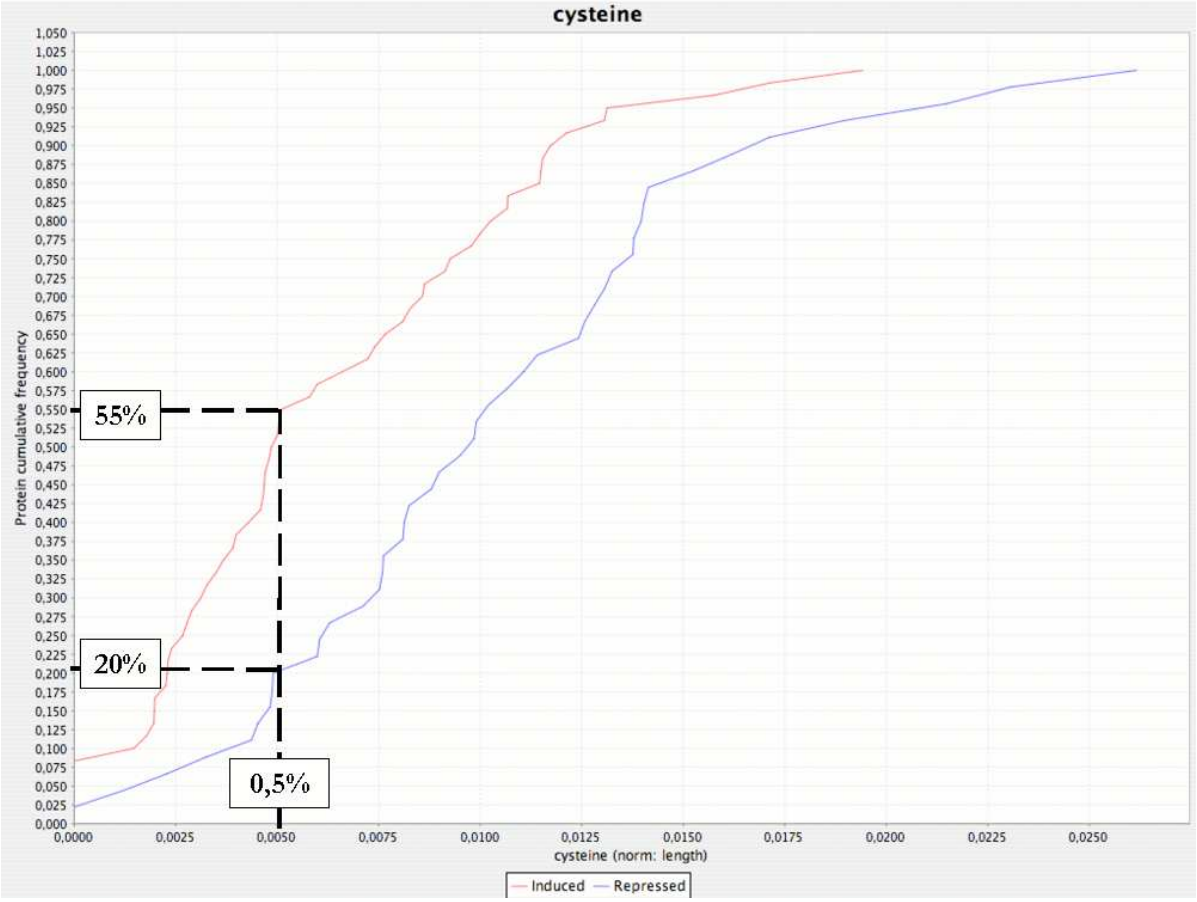


FIG. 1.9 – **Biais en cystéine chez la levure.** Ce graphique représente les fréquences cumulées des protéines possédant un pourcentage donné de cystéine [Baudouin-Cornu *et al.*, 2001]. Les abscisses représentent en effet le pourcentage d'acides aminés étant des cystéines (le nombre de cystéines a été normalisé par la longueur de la protéine). Les deux courbes montrent les fréquences cumulées pour les deux groupes considérés, c'est-à-dire les 60 protéines correspondant aux gènes induits et les 45 protéines correspondant aux gènes réprimés. Prenons un exemple pour illustrer la lecture de ce graphique : dans le cas des induits (en rouge), 55% des protéines contiennent moins de 0,5% de cystéines ; dans le cas des réprimés (en bleu), 20% des protéines contiennent moins de 0,5% de cystéines. De manière globale, ce graphique montre que les induits possèdent un plus faible pourcentage de cystéines. Notons qu'il ne permet en rien de conclure quant à l'existence d'un biais significatif.

1.4.2 Application à *Synechocystis*

Nous avons ensuite appliqué la méthode de détection automatique de biais à *Synechocystis* afin d'identifier d'éventuels biais dans la composition des protéines au cours des réponses transcriptionnelles aux stress étudiés. Les résultats détaillés sont présentés dans l'annexe C.

1.4.2.1 La cinétique Cd chez *Synechocystis*

Dans un premier temps, nous avons étudié la cinétique Cd chez *Synechocystis* en considérant les gènes globalement régulés, puis les gènes répondant en deux phases.

1.4.2.2 Comparaison des gènes globalement régulés

Nous avons comparé les groupes de protéines correspondant aux gènes globalement induits d'une part, et aux gènes globalement réprimés d'autre part. Les résultats des tests statistiques sont indiqués dans la Table C.2. Concernant la composition en acides aminés, nous avons mis en évidence que les protéines codées par les gènes induits contenaient moins d'alanine, de glycine et de valine, mais plus de glutamine. De plus, nous avons montré que ces protéines avaient tendance à être plus polaires, moins hydrophobes, plus grandes, plus lourdes et plus instables.

Ces différents biais peuvent s'expliquer facilement étant donnés les biais en acides aminés. D'une part, la glutamine est un acide aminé polaire. Comme les protéines codées par les gènes induits ont plus de glutamine, ces protéines ont davantage tendance à être polaires. D'autre part, les acides aminés alanine, glycine et valine sont hydrophobes, de petite taille et légers. Ainsi, leur faible présence chez les protéines codées par les gènes induits a pour conséquence des protéines moins hydrophobes, plus grandes et plus lourdes.

1.4.2.3 Comparaison des gènes répondant en deux phases

Nous avons comparé les groupes de protéines correspondant aux gènes induits en phase massive d'un côté, et aux gènes réprimés en phase massive de l'autre. Les résultats des tests statistiques sont indiqués dans la Table C.3. Concernant la composition en acides aminés, nous avons mis en évidence que les protéines codées par les gènes induits contenaient moins de lysine, mais plus de leucine, de glutamine et de tryptophane. De plus, nous avons montré que ces protéines avaient tendance à être moins hydrophobes, plus instables et à absorber plus la lumière.

Notons que les biais en acides aminés, dans ce cas, apportent des contributions complémentaires par rapport au caractère hydrophobe des protéines. En effet, les protéines correspondant aux gènes induits contiennent plus de leucine et de tryptophane, qui sont deux acides aminés hydrophobes, mais moins de lysine, qui est également un acide aminé hydrophobe. Ceci suggère que le biais en lysine serait quantitativement plus important que les deux autres.

De plus, nous avons montré que la machinerie photosynthétique était dégradée lors de la réponse au stress Cd [Houot *et al.*, 2007]. Or l'appareil photosynthétique est for-

tement hydrophobe. Par conséquent, ceci peut expliquer l'origine du biais en termes d'hydrophobicité entre les deux groupes de protéines.

1.4.2.4 La cinétique H_2O_2 chez *Synechocystis*

Nous avons ensuite réalisé la même étude pour la cinétique H_2O_2 chez *Synechocystis* en considérant les gènes globalement régulés, puis les gènes répondant en deux phases.

1.4.2.5 Comparaison des gènes globalement régulés

Dans la mesure où aucun gène n'avait été identifié comme globalement réprimé, nous avons abaissé le seuil de détection à 10^{-2} au lieu de 10^{-3}. Dans ce cas, 22 gènes étaient réprimés et 69 induits. Nous avons alors comparé les groupes de protéines correspondant à ces groupes de gènes. Les résultats des tests statistiques sont indiqués dans la Table C.4. Aucun biais n'a pu être mis en évidence dans ce cas.

1.4.2.6 Comparaison des gènes répondant en deux phases

Nous avons ensuite comparé les groupes de protéines correspondant aux gènes induits en phase massive d'un côté, et aux gènes réprimés en phase massive de l'autre. Les résultats des tests statistiques sont indiqués dans la Table C.5. Concernant la composition en acides aminés, nous avons mis en évidence que les protéines codées par les gènes induits contenaient plus d'isoleucine et de lysine. De plus, nous avons montré que ces protéines avaient tendance à être basiques, positivement chargées, grandes, longues et avec un fort coefficient d'extinction.

Le biais concernant la nature basique des protéines peut s'expliquer par la forte présence de lycine car c'est un acide aminé basique. Il est par contre étonnant de ne pas identifier un biais de polarité dans la mesure où tous les acides aminés basiques sont polaires. Cela dit, certains acides aminés sont polaires sans pour autant être acide ou basique, c'est le cas de la sérine par exemple. Il peut y avoir des différences fines en termes d'acides aminés, qui ne sont pas détectées séparément mais contribuent à l'équilibre global. C'est tout l'intérêt de la méthode de considérer cet équilibre global pour les différents paramètres. Pour étudier tous ces paramètres, nous avons été amenés à développer un outil de prédiction de biais.

1.4.3 Implémentation d'un outil de détection automatique de biais de composition

Nous avons développé un outil, BiasSeeker, permettant d'effectuer de manière automatique cette recherche de biais entre deux groupes de protéines ou de gènes (voir la figure 1.10). L'implémentation a été réalisée par Guillaume Meurice. Tous les paramètres ont été calculés à partir des séquences protéiques (voir le détail en Annexe B.2). Ainsi, nous avons choisi de considérer en entrée deux fichiers au format FASTA donnant les séquences des protéines de chaque groupe (voir les champs *fasta file #1* et *fasta file #2*

de la figure 1.11). Le format FASTA fournit les séquences des protéines, ainsi qu'un identifiant pour chacune des protéines.

Après que tous les tests ont été réalisés pour les différents paramètres, les résultats des différents tests statistiques sont affichés, permettant ainsi d'identifier les biais potentiels. De plus, les graphiques associés permettent de savoir dans quel sens est le biais s'il existe (voir Figure 1.12).

En définitive, nous avons développé une méthode de détection automatique de biais entre deux groupes de protéines. De plus, nous l'avons implémentée sous la forme d'un outil avec une interface graphique rendant son utilisation simple et pratique. Par ailleurs, nous avons validé notre méthode grâce à des travaux réalisés au préalable. Enfin, aucun biais en soufre n'a pu être mis en évidence chez *Synechocystis*. Même si cette étude ne permet pas de conclure à l'absence d'un mécanisme d'économie du soufre chez *Synechocystis*, cette hypothèse semble probable.

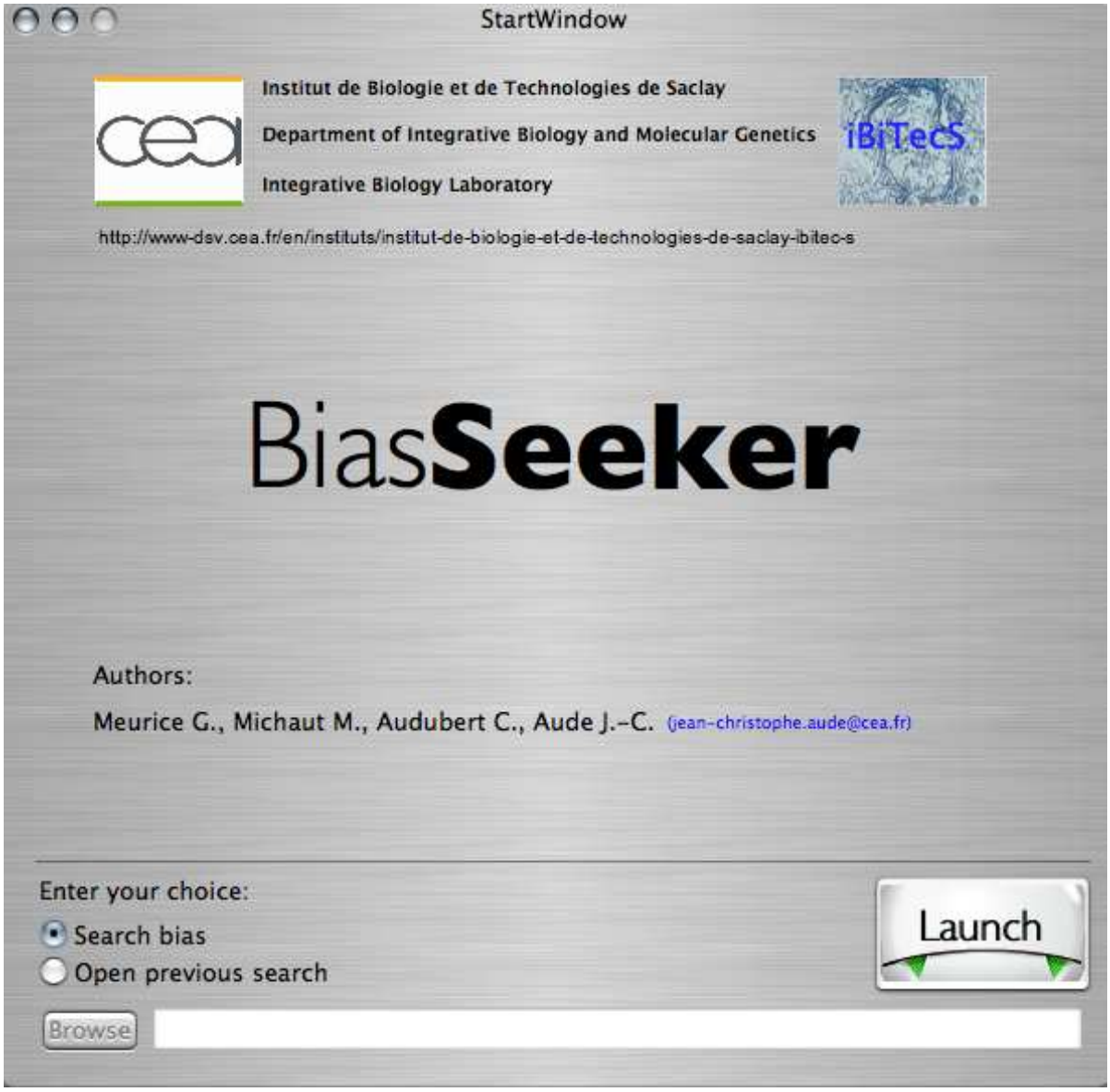


FIG. 1.10 – **Page d'accueil de BiasSeeker.** Cette image est une capture d'écran de la page d'accueil de l'outil BiasSeeker. À partir de là, il est possible de recharger des analyses antérieures ou de réaliser une nouvelle recherche de biais.

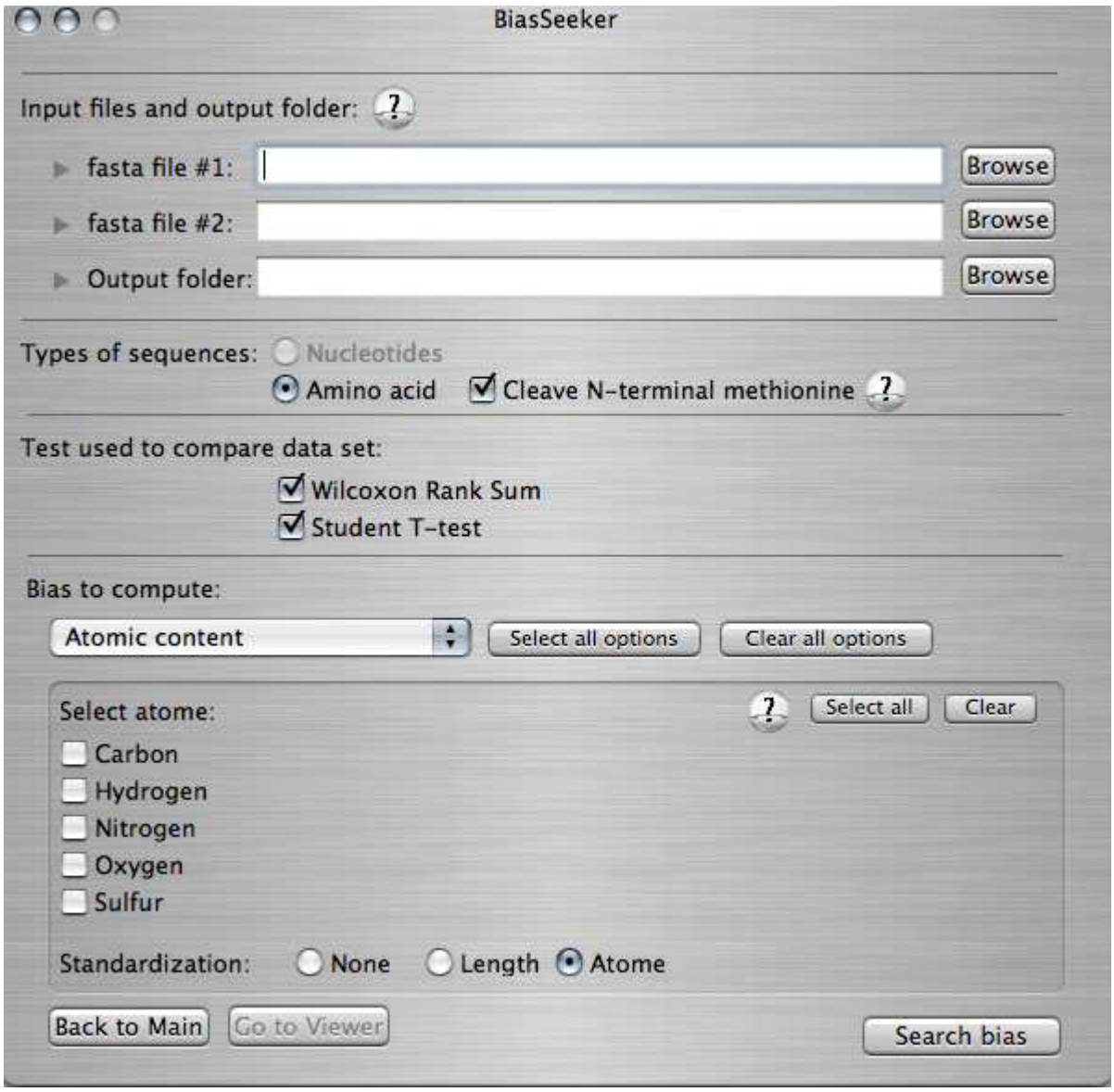


FIG. 1.11 – **Paramétrage de BiasSeeker.** Cette image est une capture d'écran de la page de paramétrage de l'outil BiasSeeker. La première partie permet à l'utilisateur de spécifier les fichiers décrivant les deux groupes de protéines (format FASTA), et le répertoire de sortie où toutes les images et les tables de valeurs sont stockées. L'utilisateur précise ensuite s'il veut considérer des acides aminés ou des nucléotides. Il a également la possibilité d'ajouter une règle de clivage de la méthionine en position N-terminale qui signale le début de la traduction (voir la Section B.2.3 de l'annexe B.1). Les tests statistiques réalisés sont ensuite précisés. Puis, les paramètres à étudier sont complétés pour chacune des quatre catégories : composition en atomes, composition en acides aminés, paramètres des acides aminés, paramètres des protéines, en précisant au besoin une normalisation en fonction de la longueur ou du nombre d'atomes.

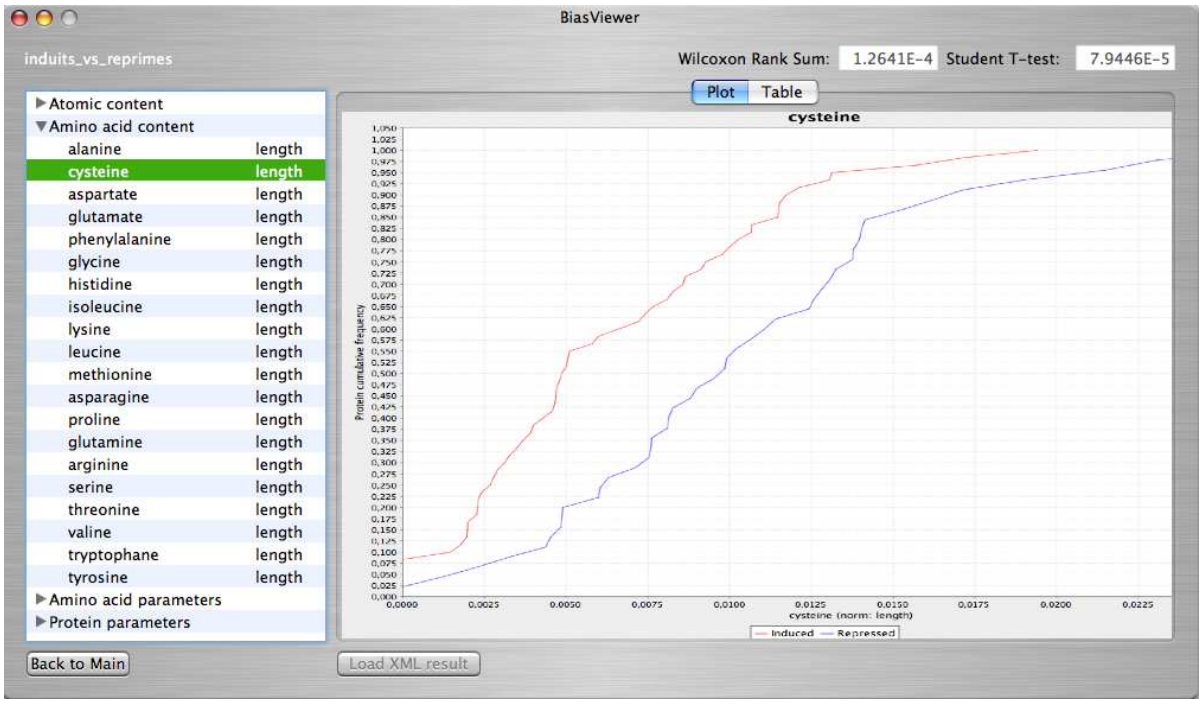


FIG. 1.12 – **Résultats de BiasSeeker.** Cette image est une capture d'écran de la page de résultats de l'outil BiasSeeker. La partie gauche permet de parcourir les résultats pour les différents paramètres étudiés, classés selon les quatre catégories suivantes : composition en atomes, composition en acides aminés, paramètres des acides aminés, paramètres des protéines. La fenêtre graphique, à droite, permet de visualiser les fréquences cumulées des protéines pour le paramètre étudié sur les deux échantillons identifiés par la légende placée en-dessous du graphique. Les résultats numériques des tests statistiques effectués sont affichés en-haut à droite. Il est également possible d'afficher la table des valeurs pour le paramètre étudié en cliquant sur l'onglet **Table**, puis de revenir à l'affichage graphique en cliquant sur l'onglet **Plot**.

Conclusion

Dans ce chapitre, nous avons montré que les réponses transcriptionnelles aux stress Cd et H_2O_2 présentent deux phases principales. Pour le Cd, nous avons identifié une première phase dite précoce, où peu de gènes sont régulés, puis une seconde phase de réponse beaucoup plus massive, où plus de 30% des gènes de *Synechocystis* sont régulés. Dans le cas de l'H_2O_2, la réponse est plus rapide. Elle consiste en une première phase massive suivie d'une phase dite tardive, où l'expression des gènes revient progressivement à son niveau habituel. Ce travail fait partie d'une étude plus générale sur le métabolisme de *Synechocystis* et le régulateur *slr1738* [Houot *et al.*, 2007] (voir l'article à la fin du manuscrit et la liste des publications page 260).

De plus, nous avons développé une méthode de classification mixte hiérarchique-pyramidale, qui nous a permis d'identifier des classes de gènes dont l'expression varie de manière similaire. Cette méthode a l'avantage d'être très générale, car elle se base sur une distance, et le découpage est automatique. Ainsi, il serait intéressant de l'appliquer sur des distances provenant de différents types d'information, comme par exemple l'annotation fonctionnelle, ou encore des distances qui combinent plusieurs informations, comme par exemple les données d'expression et la localisation. La méthode de découpage automatique pourrait également être améliorée, notamment par des approches plus globales de qualification des partitions sur l'ensemble de la hiérarchie. Ce travail a fait l'objet d'un chapitre de livre [Polaillon *et al.*, 2007] (voir l'article à la fin du manuscrit et la liste des publications page 260).

Par ailleurs, nous avons développé une méthode de détection de biais de composition entre des classes de protéines, que nous avons appliquée aux classes de gènes régulés. Ceci nous a permis de caractériser les classes de protéines correspondantes en termes de composition. La limite principale de cette méthode est que nous ne tenons compte ni des niveaux d'induction des gènes, ni de la relation quantitative entre les ARNm et les protéines. Par ailleurs, les relations entre les différents biais pourraient être analysées plus en détails. Néanmoins, la dépendance entre les différents paramètres étudiés rend difficile la réalisation de tests statistiques pertinents. Ce problème mériterait d'être traité plus en profondeur. Enfin, cette méthode de détection de biais de composition a été implémentée sous la forme de l'outil BiasSeeker qui rend son utilisation simple, rapide et visuelle. Cette étude sur les biais est encore en cours. L'outil BiasSeeker devrait être utilisé pour étudier en particulier les génomes de différentes cyanobactéries.

Après avoir identifié des classes de protéines, nous avons voulu expliciter les relations physiques entre ces protéines. À cette étape du travail, très peu de données expérimentales étaient disponibles pour *Synechocystis* (moins de 200 interactions protéine-protéine). Par conséquent, l'objectif a été de construire un réseau d'interactions protéine-protéine pour *Synechocystis*.

Chapitre 2

Inférence *in-silico* de réseaux d'interactions protéine-protéine

"Il y a là de formidables défis, mais ils deviennent quasiment insignifiants quand on considère la complexité du niveau suivant, celui des interactions entre les protéines."

Denis Noble,
La musique de la vie
La biologie au-delà du génome, 2007

À cette étape, l'objectif était de construire un réseau d'interactions protéine-protéine pour *Synechocystis*, c'est-à-dire d'établir une liste d'interactions protéine-protéine potentielles. Pour cela, nous avons adapté et développé des méthodes de prédiction d'interactions. Nous les avons ensuite appliquées à *Synechocystis* pour construire un réseau. Finalement, nous avons analysé les résultats, notamment en confrontant les interactions prédites à d'autres types de données, comme par exemple des interactions identifiées expérimentalement, des annotations fonctionnelles et des annotations de domaines d'interaction, mais aussi en identifiant certaines interactions prédites en considérant par exemple la conservation des interologues et la multiplicité des techniques expérimentales ayant permis de mettre en évidence les interactions sources. Notons pour finir que l'une des méthodes de prédiction a donné naissance à un outil automatique de prédiction d'interactions protéine-protéine.

2.1 Développement de méthodes de prédiction *in-silico* d'interactions protéine-protéine

Les méthodes de prédiction d'interactions protéine-protéine que nous avons développées sont basées sur le concept d'interologue qui combine des interactions connues avec des relations d'orthologie, afin de transférer des interactions protéine-protéine d'un organisme à un autre (voir page 54). Cette approche est considérée comme un transfert d'interactions d'une espèce dite *source* vers une espèce dite *cible* (voir Figure 2.1). Nous avons sélectionné sept organismes pour lesquels les interactions protéine-protéine ont déjà été étudiées : *Saccharomyces cerevisiae*, *Escherichia coli*, *Homo sapiens*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophilia melanogaster*, *Helicobacter pylori*. Pour déterminer les relations entre les protéines de ces espèces et celles de *Synechocystis*, nous avons utilisé l'homologie de séquences. Ainsi, les différentes méthodes de prédiction que nous avons développées se différencient principalement les unes des autres par les stratégies utilisées pour identifier les relations potentielles d'orthologies.

2.1.1 La méthode *InteroRBH*

La méthode *InteroRBH* est basée sur la combinaison d'interactions et de liens potentiels d'orthologie identifés par l'approche RBH (*Reciprocal Best Hit*) [Tatusov *et al.*, 1997] [Hirsh et Fraser, 2001] [Jordan *et al.*, 2002], d'où le nom de la méthode *InteroRBH*.

Pour toutes les protéines de l'espèce source, nous commençons par identifier les protéines potentiellement orthologues chez l'espèce cible. Pour cela, nous utilisons des comparaisons de séquences calculées avec l'algorithme de Smith-Waterman [Smith et Waterman, 1981] et disponibles dans la base de données CluSTr [Petryszak *et al.*, 2005]. Ces similarités de séquences sont calculées avec l'implémentation de Sencel appelée PARALIGN [Saebø *et al.*, 2005] et sont associées à une E-value. Pour sélectionner les meilleures similarités de séquences, nous avons pris en compte seulement les comparaisons avec une E-value inférieure à 10^{-10}. Cette valeur de seuil, couramment utilisée [Martin *et al.*, 2002] [Yu *et al.*, 2004b], est très sélective, et nous permet de ne considérer que des similarités de séquences très fortes.

Rappelons rapidement le principe de l'approche RBH [Tatusov *et al.*, 1997] [Hirsh et Fraser, 2001] [Jordan *et al.*, 2002]. Pour que deux protéines soient considérées comme potentiellement orthologues, il faut et il suffit qu'elles aient chacune la séquence la plus proche de l'autre dans l'espèce donnée. Illustrons ceci sur un exemple. Pour chaque protéine P_2 d'une espèce donnée (l'espèce P dans la Figure 2.2), nous sélectionnons la séquence la plus proche de celle de P_2 dans une espèce donnée (l'espèce U dans la Figure 2.2), c'est-à-dire celle avec la similarité de séquence la plus forte (si elle existe). Réciproquement, pour chaque protéine de l'espèce U, nous recherchons la séquence qui lui est la plus similaire dans l'espèce P. Si P_2 est la séquence la plus proche de U_3, et U_3 est la séquence la plus proche de P_2, alors on considère qu'il y a un lien potentiel d'orthologie réciproque entre ces deux protéines. P_2 est la protéine orthologue de U_3 dans l'espèce P, et U_3 est la protéine orthologue de P_2 dans l'espèce U. De cette façon,

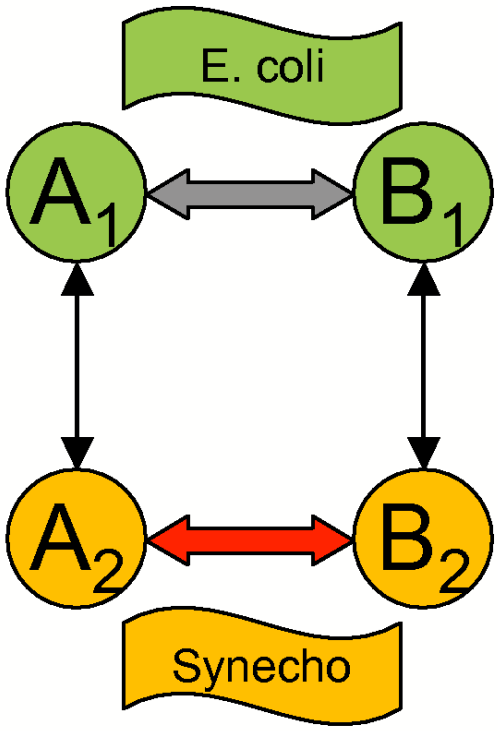


FIG. 2.1 – **Méthode d'inférence par interologue.** Nous illustrons ici le concept d'interologue sur lequel sont basées les méthodes de prédiction d'interactions protéine-protéine. Il s'agit du transfert d'interactions d'une espèce *source* (*E. coli*) vers une espèce *cible* (*Synechocystis*). Pour une interaction source donnée dans l'espèce source, nous considérons chacune des deux protéines en interaction (les protéines A_1 et B_1 chez *E. coli*), et nous recherchons des orthologues dans *Synechocystis*. Si les deux protéines ont une protéine orthologue, ceci conduit à la prédiction d'une interaction chez *Synechocystis* entre les protéines A_2 et B_2.

une protéine ne peut avoir au plus qu'une protéine potentiellement orthologue dans une espèce donnée.

Une fois que les relations d'orthologie sont identifiées, nous pouvons effectuer le transfert des interactions de l'espèce source vers l'espèce cible. Prenons par exemple *E. coli* comme espèce source et *Synechocystis* comme espèce cible (voir Figure 2.1). Pour chaque interaction décrite dans *E. coli*, nous considérons chacune des deux protéines, et nous recherchons des protéines orthologues dans *Synechocystis*. Si les deux protéines en interaction ont chacune une protéine potentiellement orthologue dans *Synechocystis*, alors nous transférons cette interaction source, ceci conduisant à la prédiction d'une interaction entre ces deux protéines de *Synechocystis*.

Les interactions prédites sont qualifiées par la E-value jointe, définie comme la moyenne géométrique des E-values de chacune des deux relations d'orthologie utilisées au cours du transfert [Yu *et al.*, 2004b].

2.1.2 La méthode *InteroBH*

L'approche *InteroBH* est tout à fait semblable à l'approche *InteroRBH* présentée précédemment. La différence réside dans la stratégie adoptée pour identifier les protéines orthologues potentielles. La réciprocité n'est plus nécessaire ici entre les deux espèces pour définir une relation d'orthologie potentielle. En d'autres termes, la protéine P_2 de la figure 2.2 aura pour protéine orthologue, dans l'espèce U, la protéine qui a la séquence la plus proche. Illustrons ceci sur un exemple.

Pour chaque protéine P_2 d'une espèce donnée (l'espèce P dans la Figure 2.2), nous sélectionnons la séquence la plus proche de celle de P_2 dans une espèce donnée (l'espèce U), c'est-à-dire celle avec la similarité de séquence la plus forte (si elle existe), et la considérons comme une protéine orthologue potentielle (la protéine U_3). De plus, si P_2 est elle-même la séquence la plus proche d'une autre protéine U_2 de cette espèce, nous ajoutons U_2 à l'ensemble des protéines potentiellement orthologues de P_2. Cette méthode ne nécessitant plus de réciprocité, nous l'appelons donc BH pour Best Hit.

Le processus d'inférence est le même que celui de la méthode InteroRBH (voir Section 2.1.1) et les interactions prédites sont également qualifiées par la E-value jointe. Si une protéine a plusieurs protéines orthologues potentielles, nous inférons alors toutes les interactions possibles.

2.1.3 La méthode *InteroPorc*

Si des transferts ont déjà été effectués pour certains organismes modèles (voir page 54), nous avons remarqué que chaque étude est faite manuellement pour l'espèce cible considérée. Nous avons alors voulu développer une méthode plus générale, utilisable pour toutes les espèces. Pour cela, nous avons utilisé les groupes de protéines potentiellement ortologues décrits dans les données PORCs (Putative ORthologous Cluster) et nous avons appelé la méthode *InteroPorc*. Celle-ci est une généralisation des approches précédentes.

Les données PORCs sont basées sur les comparaisons de séquences de la base de

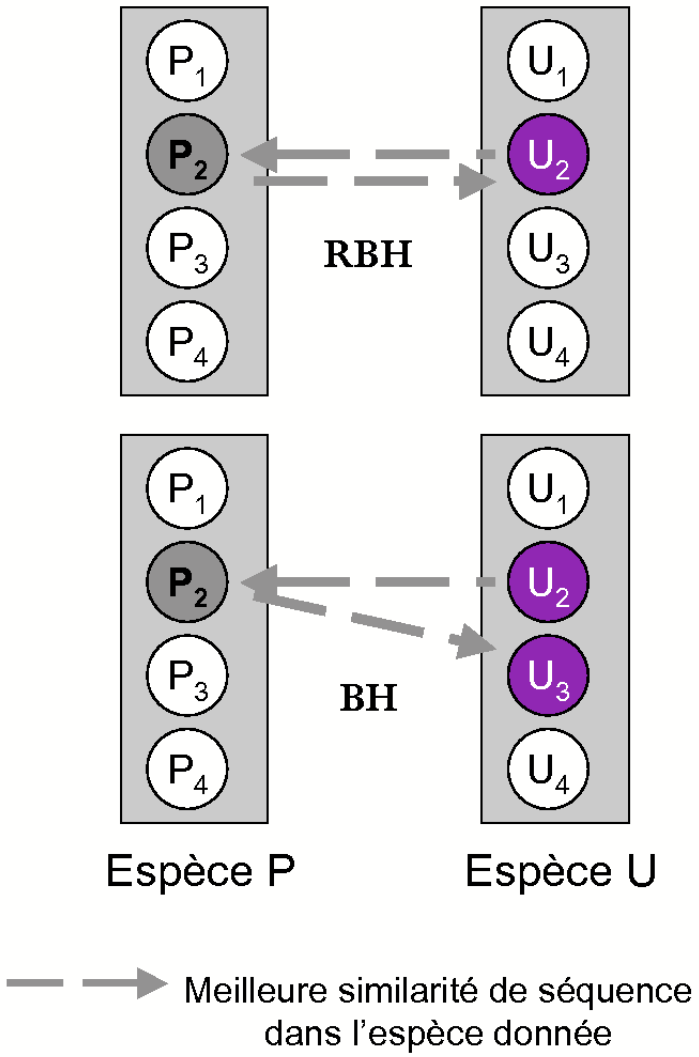


FIG. 2.2 – **Identification des protéines orthologues potentielles.** Nous montrons ici la stratégie utilisée pour identifier les protéines orthologues potentielles dans une espèce donnée. Pour chaque protéine P_2 de l'espèce P, nous sélectionnons la séquence la plus proche de celle de P_2 dans l'espèce U, c'est-à-dire celle avec la similarité de séquence la plus forte (U_2 en haut, U_3 en bas). Réciproquement, nous recherchons la protéine la plus proche dans l'espèce P. L'approche RBH (Reciprocal Best Hit) sélectionne U_2 comme orthologue potentielle de P_2 si et seulement si la relation est réciproque (cas du haut). Cette condition n'est pas nécessaire pour l'approche BH (Best Hit) pour laquelle U_2 et U_3 sont sélectionnées (cas du bas).

données CluSTr [Petryszak *et al.*, 2005] et font partie d'Integr8 [Kersey *et al.*, 2005]. La base de données Integr8 met à disposition l'ensemble des génomes séquencés, réactualisé tous les mois, ainsi que leur protéome correspondant. En janvier 2008, Integr8 contenait 655 organismes cellulaires différents dont 556 bactéries, 59 eucaryotes et 50 archées. Quelques mois plus tard, en octobre 2008, la base contenait 785 organismes cellulaires et 501 bactériophages. Ces groupes de protéines sont donc d'un très grand intérêt puisqu'ils contiennent tous les organismes dont le génome est séquencé. Ceci n'est pas le cas pour la principale classification du même type que sont les COGs [Tatusov *et al.*, 1997].

Chaque élément des données PORC représente un cluster de gènes regroupés d'après la similarité de leur produit le plus long. Nous avons utilisé 215 733 clusters contenant en tout 1 548 235 protéines. D'après le processus de construction, chaque cluster contient au plus une protéine d'une espèce donnée, et chaque protéine est assignée à au plus un cluster. En d'autres termes, il est impossible de trouver plusieurs protéines de la même espèce à l'intérieur d'un cluster. Si les comparaisons de séquences ne sont pas totalement suffisantes pour effacer toute ambiguïté, l'algorithme de construction des clusters n'utilise pas de moyens supplémentaires tels que des arbres phylogénétiques ou la comparaison de réseaux [Bandyopadhyay *et al.*, 2006]. Les clusters sont séparés en fonction du lien de similarité le plus faible pour respecter la règle selon laquelle chaque cluster contient au plus une protéine d'une espèce donnée [Kersey *et al.*, 2005].

Maintenant que nous avons décrit l'obtention de relations potentielles entre les espèces, nous allons voir comment nous les avons utilisées pour prédire des interactions protéine-protéine. Le processus d'inférence d'interactions protéine-protéine est constitué de deux étapes au cours desquelles les interactions sont abstraites sous forme d'interactions entre clusters, puis projetées pour redescendre au niveau des protéines mais chez l'espèce cible et non plus l'espèce source.

Dans un premier temps, nous utilisons les interactions sources et les clusters pour construire des liens entre les clusters (voir Figure 2.3). Pour chaque interaction source, si les deux protéines appartiennent à un cluster, nous construisons alors un lien entre ces deux clusters. Ce faisant, le niveau d'abstraction change pour passer des protéines et des interactions protéine-protéine au niveau des clusters et des liens entre clusters. Cette étape est ainsi appelée up-casting en référence au travail de thèse de M. Lappe [Lappe, 2003].

Dans un second temps, nous utilisons des interactions entre clusters pour prédire des interactions protéine-protéine chez l'espèce cible. Ainsi, pour une interaction donnée entre deux clusters, si chacun des deux clusters contient une protéine de l'organisme considéré, nous prédisons alors l'existence d'une interaction protéine-protéine entre ces deux protéines. Cette étape est appelée down-casting [Lappe, 2003].

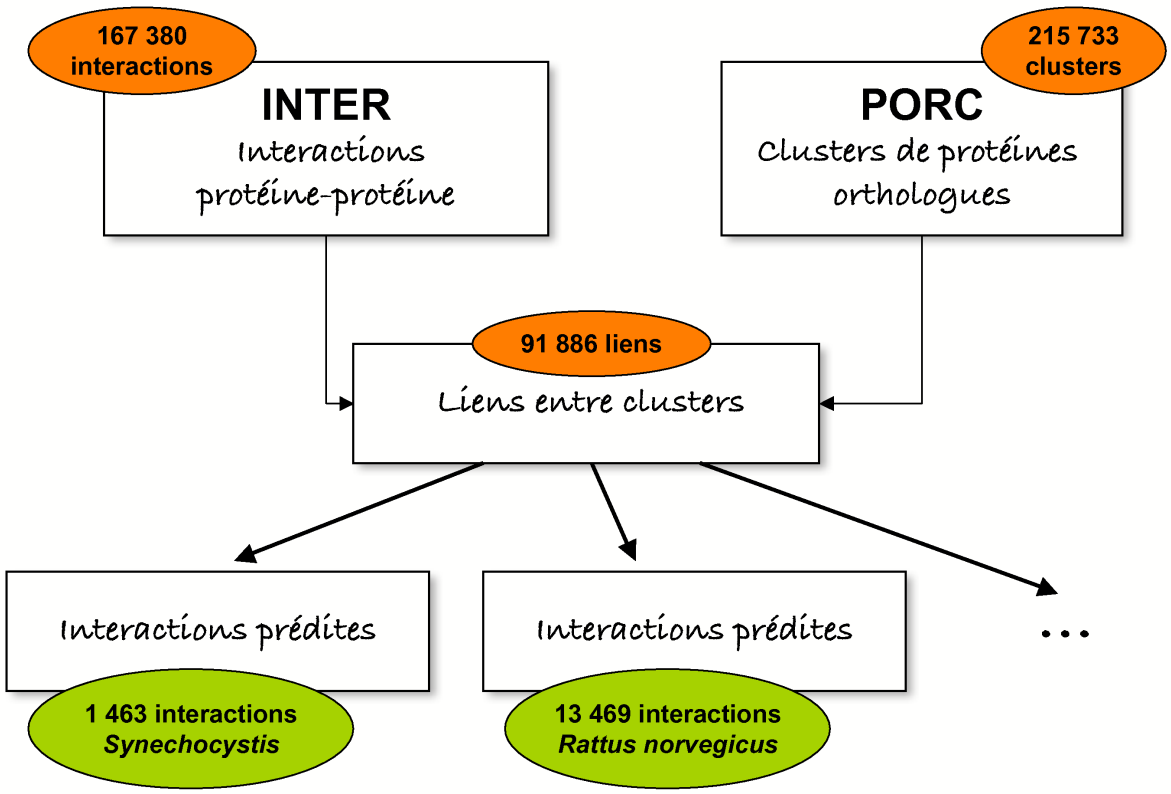


FIG. 2.3 – **Illustration de la méthode InteroPorc.** Nous disposons, d'une part, d'interactions protéine-protéine sources (INTER sur le graphique), et, d'autre part, de clusters de protéines orthologues (PORC sur le graphique). Dans un premier temps, nous combinons ces deux informations de manière à construire des liens entre deux clusters de protéines. Dans un second temps, nous projetons ces liens sur une ou plusieurs espèces cibles (par exemple *Synechocystis* et le rat sur ce graphique).

À ce stade, nous avions développé des méthodes de prédiction d'interactions protéine-protéine, basées sur le concept des interologues, qui combinent des interactions sources et des relations d'orthologies potentielles, afin de transférer les interactions connues d'un organisme vers un autre.
Nous avons alors appliqué ces méthodes de prédiction *in-silico* de manière à construire un réseau d'interactions protéine-protéine pour l'espèce cible qui nous intéresse en tout premier lieu, à savoir la cyanobactérie *Synechocystis*.

2.2 Construction d'un réseau d'interactions protéine-protéine chez *Synechocystis*

L'objectif était d'appliquer ces méthodes de prédiction afin de construire un réseau d'interactions protéine-protéine chez *Synechocystis*. Pour cela, nous avions besoin d'interactions connues dans des organismes sources. Nous avons donc commencé par sélectionner un jeu de données d'interactions sources qui seront les mêmes pour toutes les méthodes. Nous avons ensuite construit les relations potentielles d'orthologie pour chacune des méthodes. Enfin, nous avons inféré différentes interactions protéine-protéine en appliquant les méthodes de prédiction *in-silico*.

2.2.1 Les interactions sources

Nous avons collecté les génomes et les protéomes des sept espèces considérées en utilisant la base de données Integr8 [Kersey *et al.*, 2005] (*Saccharomyces cerevisiae*, *Escherichia coli*, *Homo sapiens*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophilia melanogaster*, *Helicobacter pylori*). Pour *Escherichia coli* et *Helicobacter pylori*, plusieurs génomes de différentes souches sont séquencés. Toutefois, pour notre projet, il était important d'avoir une référence unique pour une espèce donnée. En effet, certaines interactions protéine-protéine sont reportées pour une espèce donnée sans préciser de souche spécifique. Pour pouvoir prendre en compte ces interactions, il nous fallait donc considérer un représentant unique pour chaque espèce. Pour cela, nous avons collecté les différents protéomes séquencés et regroupé les séquences très proches de manière à obtenir un méta-protéome où chaque souche est représentée. De plus, pour les protéines ayant différentes versions provenant de l'épissage alternatif, nous avons considéré seulement le produit le plus long des gènes qui les codent.

Nous avons téléchargé les jeux de données expérimentaux des trois bases de données suivantes : IntAct [Kerrien *et al.*, 2007a], MINT [Zanzoni *et al.*, 2002] et DIP [Salwinski *et al.*, 2004]. Ces bases de données rendent compte des interactions moléculaires grâce à une curation manuelle de la littérature. Elles sont acutellement les seules à fournir leurs jeux de données au format standard MITAB défini par le groupe Proteomics Standard Initiative (PSI) [Hermjakob *et al.*, 2004a] [Kerrien *et al.*, 2007b] de l'organisation travaillant sur le protéome humain (HUPO)[1]. Il s'agit d'un format texte-tabulé, où chaque ligne décrit une interaction et la manière dont elle a été identifiée. Nous avons

[1]Human Proteome Organization http ://www.hupo.org/

regroupé les trois jeux de données et supprimé les redondances de manière à ce qu'une interaction entre deux protéines données n'apparaisse qu'une seule fois. Nous n'avons considéré que les interactions physiques pour laisser de côté les interactions génétiques et les relations de colocalisation. Enfin, nous avons extrait les interactions dont les deux protéines provenaient d'une même espèce parmi celles considérées. Les 139 325 interactions sources utilisées sont indiquées dans la Table 2.1.

2.2.2 Les relations d'orthologie potentielles

Nous avons appliqué les stratégies d'identification de protéines orthologues définies précédemment, en utilisant *Synechocystis* comme espèce cible et les sept espèces sélectionnées comme organismes sources. Pour chaque protéine, nous avons calculé le nombre d'espèces dans lesquelles des protéines orthologues potentielles ont pu être identifiées parmi les sept espèces sélectionnées (voir Figure 2.4). Nous remarquons qu'un plus grand nombre de protéines ont au moins une protéine orthologue dans une espèce avec la méthode PORC. D'un autre côté, moins de protéines ont, avec cette méthode, des protéines orthologues dans plusieurs espèces. Ceci peut s'expliquer par le fait que les espèces sélectionnées sont distantes sur le plan de l'évolution, non seulement de *Synechocystis*, mais aussi entre elles. Par conséquent, seules les protéines largement conservées peuvent être dans un cluster contenant plusieurs de ces espèces. La méthode RBH identifie un peu moins de protéines orthologues potentielles que la méthode BH, du fait de la contrainte sur la réciprocité, mais les ordres de grandeur sont très proches.

2.2.3 Les interactions obtenues avec *InteroRBH*

Nous combinons ici les données d'interactions sources avec les données d'orthologie obtenues par la méthode RBH pour transférer ces interactions, de façon indépendante, depuis les sept espèces sélectionnées vers l'espèce cible *Synechocystis*. Les résultats obtenus pour chacun des transferts sont indiqués dans la Table 2.2. En regroupant les résultats provenant des différentes espèces, nous obtenons un ensemble global de 3 495 interactions entre 726 protéines (21% du protéome de *Synechocystis*).

Le réseau global est appelé *InteroRBH_LOW*. Comme cela a été explicité précédemment, nous utilisons la E-value jointe définie par Yu *et al.* pour qualifier les interactions prédites (voir section 2.1.1) [Yu *et al.*, 2004b]. Comme toutes les comparaisons de séquences ont une E-value inférieure à 10^{-10}, la E-value jointe de chaque interologue est donc également inférieure à 10^{-10}. Nous allons extraire de ce réseau global deux sous-réseaux imbriqués l'un dans l'autre.

D'abord, il a été montré qu'un seuil de 10^{-70} sur la E-value jointe permet de transférer des interactions protéine-protéine avec une grande confiance [Yu *et al.*, 2004b]. Par conséquent, nous avons considéré tous les interologues avec une E-value inférieure à 10^{-70} comme un jeu de données particulier, appelé *InteroRBH_HIGH*, et pour lequel les homologies de séquences sont les plus fortes.

Ensuite, il faut noter que les comparaisons de séquences utilisées au cours du processus de construction des PORCs sont filtrées avec un seuil à 10^{-40}. Ainsi, dans un

Organismes	Interactions	Protéines
S. cerevisiae	54 560	5 780
A. thaliana	1 406	758
E. coli	22 023	3 853
H. sapiens	26 587	9 234
D. melanogaster	27 476	8 636
C. elegans	5 636	3 275
H. pylori	1 637	783

TAB. 2.1 – **Interactions protéine-protéine des organismes sources.** Pour chaque organisme sont indiqués le nombre d'interactions collectées à partir des trois bases de données IntAct, MINT et DIP, ainsi que le nombre de protéines en interaction.

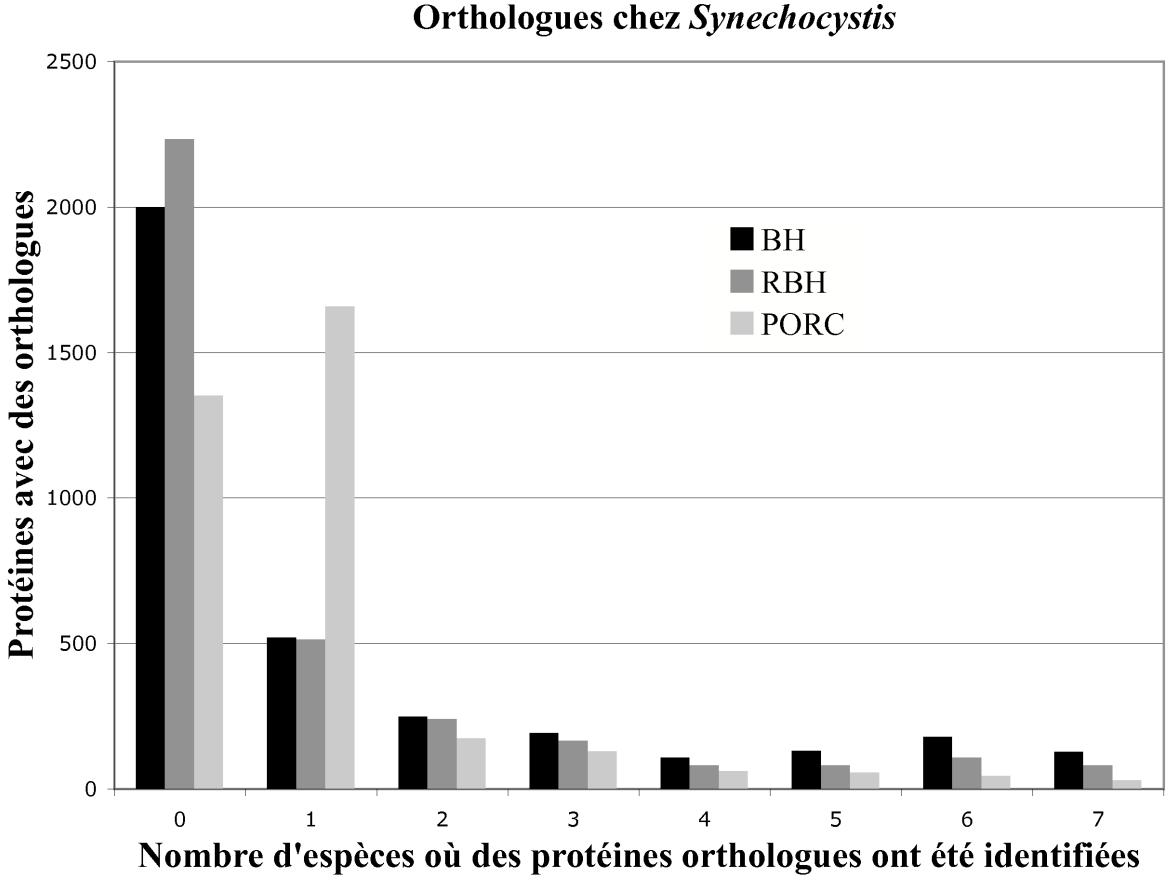


FIG. 2.4 – **Identification des protéines potentiellement orthologues chez *Synechocystis*.** Pour chaque protéine de *Synechocystis*, nous avons calculé le nombre d'espèces dans lesquelles des protéines orthologues potentielles ont pu être identifiées parmi les sept espèces sélectionnées pour cette étude. Ce nombre dépend de la stratégie utilisée pour définir les protéines orthologues potentielles : BH pour Best Hit, RBH pour Reciprocal Best Hit et PORC pour Putative ORthologous Clusters.

souci de cohérence, nous avons construit un jeu de données intermédiaire, appelé *InteroRBH_MEDIUM*, dont tous les interologues ont une E-value jointe inférieure à 10^{-40}.

En déplaçant le seuil sur la E-value considérée, nous avons donc défini trois réseaux *InteroRBH_LOW*, *InteroRBH_MEDIUM* et *InteroRBH_HIGH*, avec des homologies de séquences de plus en plus fortes.

Sans surprise, nous avons remarqué que le nombre d'interactions prédites dépend du nombre d'interactions disponibles pour l'espèce source (voir les Tables 2.1 et 2.2). Cela dépend également de la proximité avec *Synechocystis* en termes d'évolution. En effet, avec presque le même nombre d'interactions sources, nous avons transféré en moyenne plus de 400 fois plus d'interactions entre la bactérie *E. coli* et *Synechocystis* qu'entre *H. sapiens* et *Synechocystis* qui appartiennent à deux règnes différents. Ceci confirme les résultats récents de Brown et Jurisica qui ont montré que le nombre d'interactions prédites par le concept des interologues dépend de la distance taxonomique entre les organismes étudiés [Brown et Jurisica, 2007].

2.2.4 Les interactions obtenues avec *InteroBH*

Nous combinons ici les données d'interactions sources avec les données d'orthologie obtenues par la méthode BH pour transférer ces interactions, de façon indépendante, depuis les sept espèces sélectionnées vers l'espèce cible *Synechocystis*. Les résultats obtenus pour chacun des transferts sont indiqués dans la Table 2.3. En regroupant les résultats provenant des différentes espèces, nous obtenons un ensemble global de 8 586 interactions entre 998 protéines (28% du protéome de *Synechocystis*). Nous utilisons la même approche que précédemment (voir Section 2.2.3) pour définir les trois réseaux *InteroBH_LOW*, *InteroBH_MEDIUM* et *InteroBH_HIGH*, avec des homologies de séquences de plus en plus fortes.

De même que pour la méthode *InteroRBH*, nous notons ici que le nombre d'interactions prédites dépend du nombre d'interactions sources, mais aussi beaucoup de la proximité en termes d'évolution entre *Synechocystis* et l'organisme source [Brown et Jurisica, 2007]. Nous avons transféré en moyenne plus de 40 fois plus d'interactions entre la bactérie *E. coli* et *Synechocystis* qu'entre *H. sapiens* et *Synechocystis*.

Nous avons choisi de garder le critère BH pour définir les orthologues potentiels car il est plus large que le critère RBH. Ainsi, dans la suite, nous garderons les interactions prédites par la méthode *InteroBH*.

2.2.5 Les interactions obtenues avec *InteroPorc*

En utilisant la méthode *InteroPorc*, nous avons inféré l'existence de 1 446 interactions protéine-protéine entre 384 protéines chez *Synechocystis*. Au cours de ce transfert, plus de 1 446 interactions sources ont été utilisées, indiquant que certains liens construits entre les clusters de protéines orthologues ont été créés indépendamment par plusieurs interactions sources. Étant donné que chaque cluster contient au plus une protéine d'une espèce donnée, ces interactions sources proviennent d'espèces différentes. Ceci met donc en évidence des interactions conservées à travers plusieurs espèces.

Organismes	**Inter-H**	**Prot-H**	**Inter-M**	**Prot-M**	**Inter-L**	**Prot-L**
S. cerevisiae	201	118	313	152	446	205
A. thaliana	0	0	0	0	0	0
E. coli	1 237	472	2 137	572	2 908	627
H. sapiens	2	4	5	10	7	13
D. melanogaster	1	2	6	11	10	18
C. elegans	0	0	0	0	1	2
H. pylori	71	60	111	92	159	126

TAB. 2.2 – **Nombre d'interactions protéine-protéine prédites chez *Synechocystis* par *InteroRBH*** Nous considérons ici les trois jeux de données suivants : *InteroRBH_HIGH* (H), *InteroRBH_MEDIUM* (M) et *InteroRBH_LOW* (L). Pour chaque espèce source sont indiqués le nombre d'interactions prédites et le nombre de protéines impliquées dans ces interactions.

Organismes	**Inter-H**	**Prot-H**	**Inter-M**	**Prot-M**	**Inter-L**	**Prot-L**
S. cerevisiae	955	299	1 826	360	3 558	438
A. thaliana	0	0	5	7	10	11
E. coli	1 775	613	3 183	744	4 894	825
H. sapiens	26	26	69	74	194	150
D. melanogaster	14	16	30	37	97	95
C. elegans	1	2	3	6	21	35
H. pylori	99	75	164	117	251	160

TAB. 2.3 – **Nombre d'interactions protéine-protéine prédites chez *Synechocystis* par *InteroBH*.** Nous considérons ici les trois jeux de données suivants : *InteroBH_HIGH* (H), *InteroBH_MEDIUM* (M) et *InteroBH_LOW* (L). Pour chaque espèce source sont indiqués le nombre d'interactions prédites et le nombre de protéines impliquées dans ces interactions.

Nous présentons ici les résultats obtenus à partir du même ensemble d'interactions sources utilisé pour les autres méthodes dans un souci de cohérence. Il faut malgré tout noter que la méthode *InteroPorc* peut s'appliquer facilement à toutes les espèces présentes dans les bases de données sources. Nous avons également réalisé les prédictions en considérant toutes les interactions sources possibles. Nous avons alors obtenu 1 463 interactions prédites chez *Synechocystis* (voir Table 2.5), seulement 1% de plus que lorsque nous nous étions basés sur les sept espèces qui possèdent le plus d'interactions connues.

En appliquant les méthodes de prédiction *in-silico* que nous avons développées, *InteroPorc* et *InteroBH_LOW*, nous avons construit un réseau global de 8 783 interactions protéine-protéine chez *Synechocystis*. Ce réseau global sera dénommé *InteroFull* dans la suite.
Nous avons alors voulu analyser ces interactions prédites en les confrontant à d'autres types d'information.

2.3 Analyse du réseau d'interactions protéine-protéine chez *Synechocystis*

Dans le but de soutenir certaines interactions prédites, nous avons exploré différentes approches basées sur les domaines d'interaction, les annotations fonctionnelles, la conservation à travers les organismes, les techniques de détection utilisées pour identifier les interactions sources, ou encore la mise en évidence expérimentale de certaines interactions.

2.3.1 Domaines d'interaction

Les protéines ont une structure tri-dimensionnelle dans l'espace qui est primordiale pour les interactions qui se créent entre elles. Cette structure, appelée structure tertiaire, est composée de blocs élémentaires, les domaines structuraux, qui sont considérés comme les éléments de base des protéines. Ces domaines jouent un rôle important dans les interactions entre deux protéines. Ces blocs élémentaires forment des interactions entre eux qualifiées d'interactions domaine-domaine.

Nous avons utilisé les domaines d'interaction pour tenter d'expliquer certaines interactions protéine-protéine prédites. Pour cela, nous avons collecté la composition en domaines *Pfam*[2] des protéines de *Synechocystis* à partir de la base de données Uniprot [Wu *et al.*, 2006a]. Cette base de donnée Pfam identifie les domaines structuraux qui apparaissent dans différentes familles de protéines, en utilisant des alignements multiples de séquences et des modèles de markov. Le sous-projet iPfam [Finn *et al.*, 2005] se focalise sur les interactions domaine-domaine. Il s'agit d'une ressource décrivant les interactions domaine-domaine observées dans la base de données PDB[3] qui regroupe les

[2]Pfam http ://www.sanger.ac.uk/Software/Pfam/
[3]PDB http ://www.rcsb.org/pdb/home/home.do

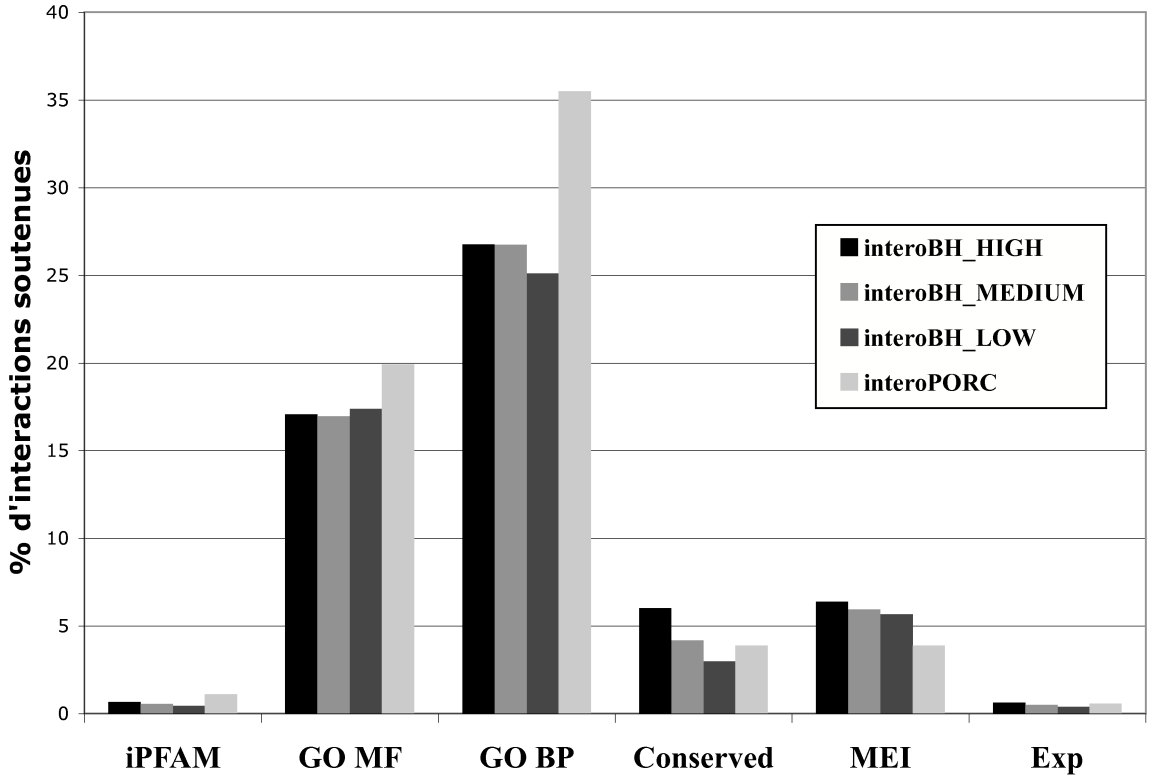


FIG. 2.5 – **Pourcentage d'interactions soutenues pour chaque réseau prédit.** Pour chaque réseau est indiqué le pourcentage d'interactions soutenues par les différentes méthodes (iPFAM : interactions expliquées par des interactions domaine-domaine, GO MF : interactions avec des termes de l'ontologie Molecular Function en commun entre les protéines en interaction, GO BP : interactions avec des termes de l'ontologie Biological Process en commun entre les protéines en interaction, Conserved : interactions prédites par différentes espèces sources, MEI : interactions prédites par des interactions identifées par plusieurs techniques expérimentales (Multiple Experimental Identification), Exp : interactions identifées expérimentalement chez *Synechocystis*).

Prédictions	**High**	**Medium**	**Low**	**Porc**
PPIs total	2 748	5 070	8 586	1 446
PPIs avec des domaines	2 689	4 939	8 197	1 399
PPIs associées à des DDIs	60	100	172	37
PPIs avec un score pertinent	18	27	38	16

TAB. 2.4 – **PPIs prédites expliquées par des interactions domaine-domaine.** L'annotation sur les interactions domaine-domaine (DDI) est décrite pour *InteroBH_HIGH* (High), *InteroBH_MEDIUM* (Medium), *InteroBH_LOW* (Lpw) et *InteroPorc* (Porc) en termes d'interactions protéine-protéine (PPI).

structures des protéines. Nous avons extrait une liste d'interactions domaine-domaine de cette ressource iPFAM. Nous avons combiné ces interactions domaine-domaine avec la composition en domaines des protéines de *Synechocystis*, de manière à extraire les interactions prédites pour lesquelles les deux protéines en interaction contiennent des domaines connus pour interagir, ceci pouvant expliquer l'interaction protéine-protéine. Pour le réseau *InteroFull* comprenant 8 783 interactions, nous avons obtenu 177 interactions dont les partenaires en interaction partagent une paire de domaines décrits comme interagissant (voir Table 2.4 pour le détail dans chacun des jeux de données).

Pour chacune des ces interactions protéine-protéine, nous avons calculé un score qui reflète la fréquence d'apparition des domaines en question, d'après l'équation 2.1.

$$S_p = \max_{d \in D}\left(\frac{1}{count(d)}\right) \tag{2.1}$$

où D est l'ensemble des interactions domaine-domaine qui peuvent être associées à l'interaction protéine-protéine p que l'on considère, et $count(d)$ est le nombre d'occurrences de l'interaction domaine-domaine d parmi toutes les paires de protéines de *Synechocystis*. Nous avons généré un ensemble de 5 000 interactions protéine-protéine aléatoirement. Dans la mesure où 95% des scores obtenus étaient en-dessous de 0,5 nous avons considéré que tous les scores au-dessus de ce seuil étaient significatifs. Par conséquent, nous considérons qu'une interaction protéine-protéine est potentiellement expliquée par une ou plusieurs interactions domaine-domaine dès que son score est au-dessus de ce seuil. Parmi les 177 interactions expliquées par des interactions domaine-domaine, 39 étaient associées à des scores qualifiés de pertinents.

Itzhaki *et al.* ont montré que les interactions domaine-domaine ont souvent lieu dans les complexes protéiques et sont conservées au cours de l'évolution [Itzhaki *et al.*, 2006]. En effet, nous observons des sous-graphes connectés qui représentent des complexes tels que le ribosome, l'ARN polymérase et l'ATP synthase (voir Figure 2.6). De plus, Itzhaki *et al.* ont trouvé que le nombre d'interactions protéine-protéine expliquées par des interactions domaine-domaine dans différents réseaux d'interactions protéine-protéine variait entre 6% et 20% seulement. Par conséquent, si on peut considérer que les interactions protéine-protéine expliquées par une ou plusieurs interactions domaine-domaine sont renforcées, on ne peut pas en conclure que celles qui ne sont associées à aucune interaction domaine-domaine en sont pénalisées. Notons que dans notre cas la valeur est d'environ 2% pour chacun des réseaux considérés (voir Table 2.4 et Figure 2.5).

2.3.2 Annotations fonctionnelles

Puisque les protéines en interaction partagent des fonctions similaires ou interviennent dans un même processus biologique [Huang *et al.*, 2007b], nous avons considéré les deux ontologies *Molecular Function* (MF) et *Biological Process* (BP) de Gene Ontology (GO) [Ashburner *et al.*, 2000] [Harris *et al.*, 2004].

Dans un premier temps, nous avons calculé le nombre de termes en commun entre les deux protéines de chaque interaction pour chacune des deux ontologies. Pour l'ontologie MF, tous les réseaux contiennent environ 30% des interactions avec au moins

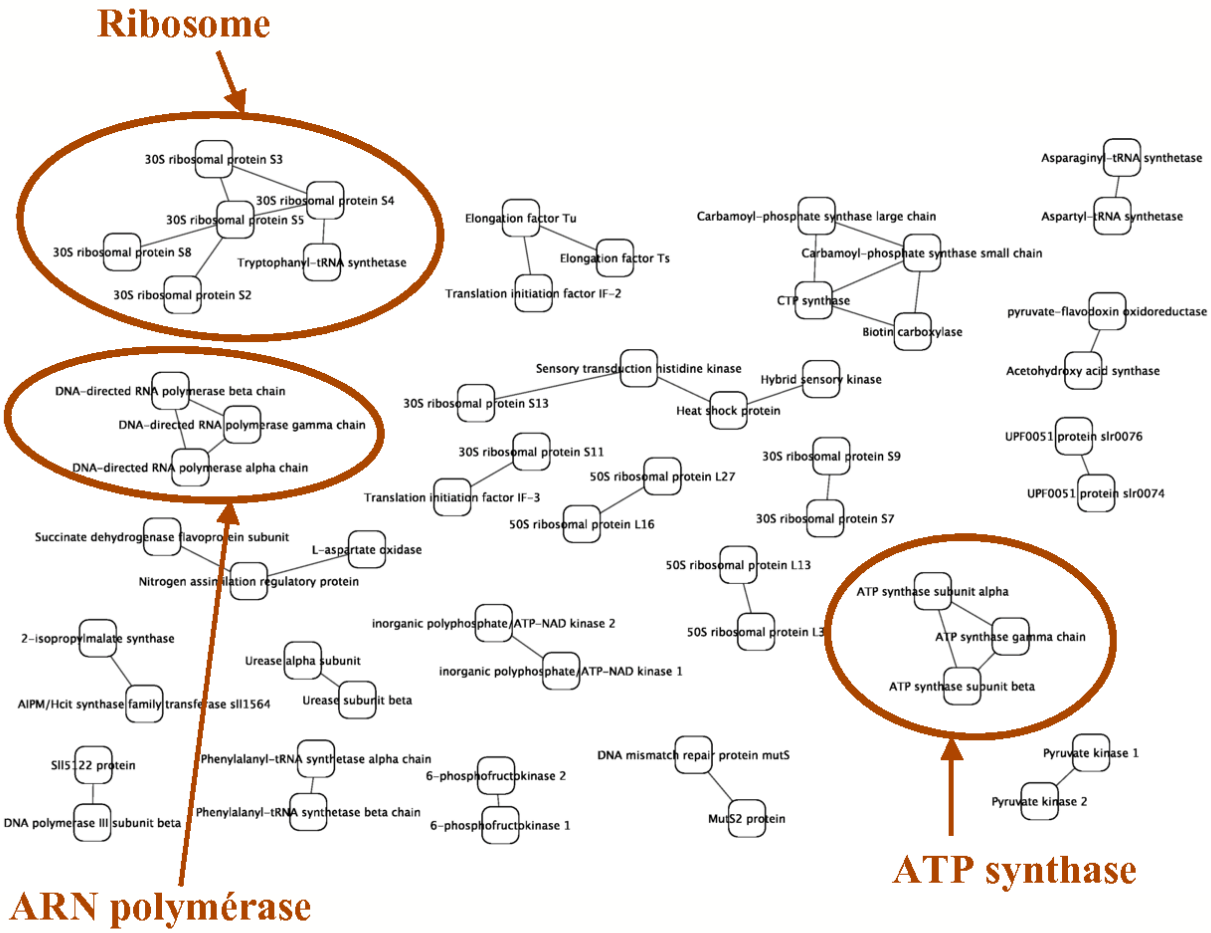


FIG. 2.6 – **Interactions soutenues par des interactions domaine-domaine.** Ce réseau représente les 39 interactions (entre 56 protéines) qui ont été soutenues par des interactions domaine-domaine avec un score pertinent. Ce graphe a été généré grâce au logiciel Cytoscape [Shannon *et al.*, 2003]. Nous observons des complexes tels que le ribosome, l'ARN polymérase et l'ATP synthase.

un terme GO en commun entre les protéines en interaction. Le réseau *InteroBH_HIGH* montre une valeur légèrement plus élevée avec 35%. Quant à l'ontologie BP, les réseaux construits avec l'approche *InteroBH* ont environ 12% des interactions ayant au moins un terme GO en commun, alors que le réseau construit par la méthode *InteroPorc* a une valeur légèrement supérieure à 17%. Si on considère le sous-graphe constitué des ces interactions, on remarque deux composantes indépendantes et largement connectées, représentant le ribosome et l'ARN polymérase (voir Figure 2.7).

Dans un second temps, nous avons voulu préciser cette approche, notamment en tenant compte de la structure en arbre de l'ontologie, et donc des liens entre les différents termes GO. Nous avons donc décidé de calculer une similarité fonctionnelle entre deux protéines, c'est-à-dire de quantifier la similarité des fonctions moléculaires ou des processus biologiques dans lesquels ces deux protéines sont impliquées. Pour cela, nous avons considéré les termes GO qui sont associés à chacune des deux protéines, et nous avons utilisé la mesure de similarité définie dans [Lubovac *et al.*, 2006]. Cette mesure est définie comme la moyenne des mesures de similarité entre les ensembles de termes GO [Lubovac *et al.*, 2006] :

$$ss(p_k, p_l) = \frac{1}{m \times n} \sum_{t_i \in T_i, t_j \in T_j} sim(t_i, t_j) \tag{2.2}$$

où

- T_k est l'ensemble des m termes de l'ontologie GO associés à la protéine p_k
- T_l est l'ensemble des n termes de l'ontologie GO associés à la protéine p_l
- $sim(t_i, t_j)$ est la mesure de similarité terme à terme définie par Lin *et al.* [Lin, 1998] et rappelée dans l'Équation 2.3 :

$$sim(t_i, t_j) = \frac{2ln\{p_{ms}(t_i, t_j)\}}{ln\{p(t_i)\} + ln\{p(t_j)\}} \tag{2.3}$$

où

- $p(t_i)$ est la probabilité du terme t_i
- $p_{ms}(t_i, t_j)$ est la probabilité du plus petit "subsumer" de t_i et t_j. Cette probabilité est définie comme la probabilité la plus faible trouvée parmi les termes parents partagés par t_i et t_j

Nous avons alors généré un ensemble de 5 000 interactions protéine-protéine aléatoirement. Dans la mesure où 95% des scores obtenus étaient en-dessous de 0,23 pour chacune des deux ontologies, nous avons considéré que tous les scores au-dessus de ce seuil étaient significatifs.

Pour l'ontologie MF, les réseaux provenant de la méthode *InteroBH* contiennent 17% d'interactions avec des annotations fonctionnelles très proches entre les deux protéines (voir Figure 2.5). Le réseau *InteroPorc* montre un pourcentage légèrement plus élevé.

Pour l'ontologie BP, les réseaux provenant de la méthode *InteroBH* ont environ 26% d'interactions entre des protéines fonctionnellement proches, comparé à 35% pour le réseau *InteroPorc*.

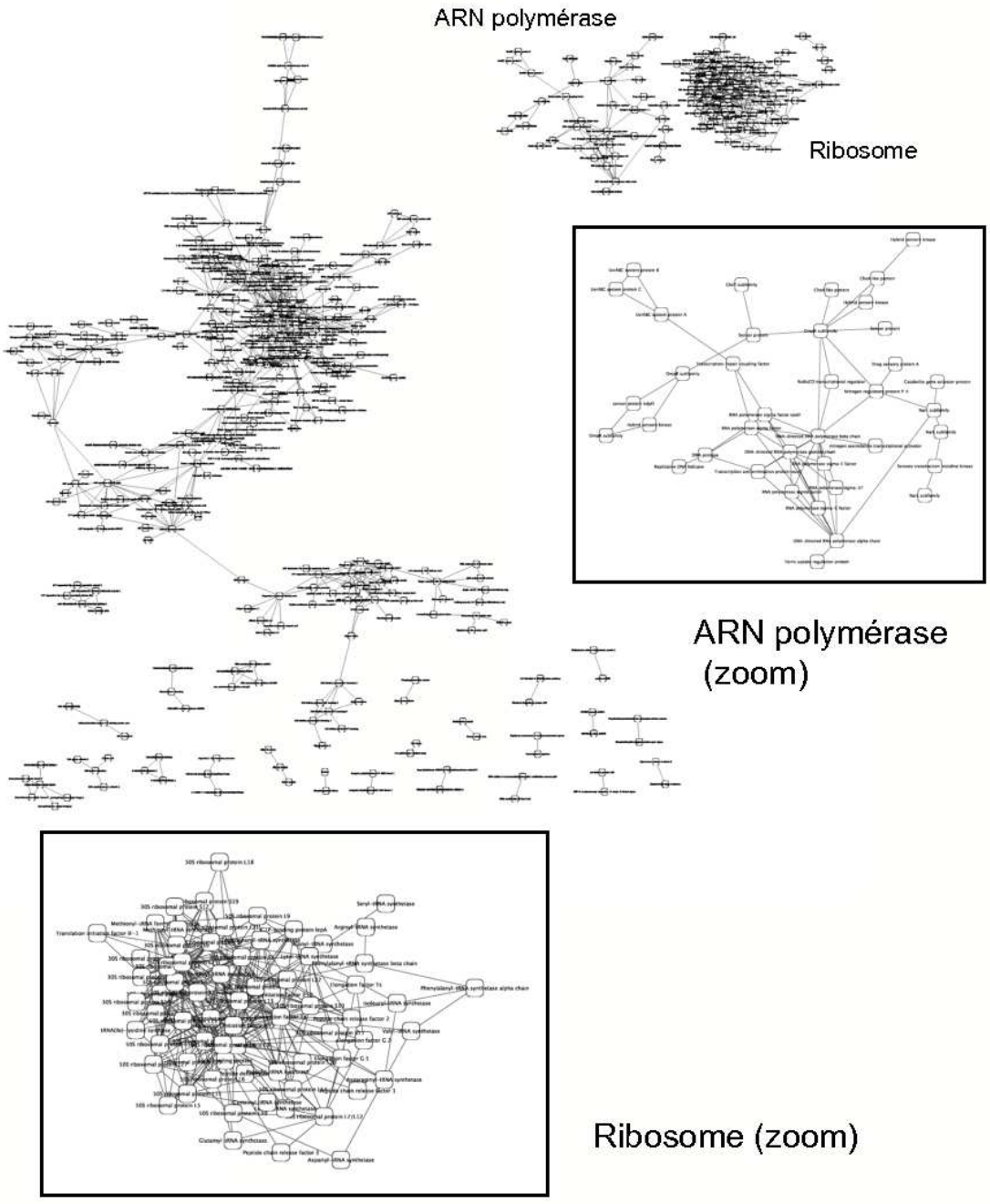


FIG. 2.7 – **Interactions soutenues par des annotations fonctionnelles.** Ce réseau représente les 981 interactions (entre 434 protéines) qui partagent des termes communs de l'ontologie 'Biological Process' entre leurs protéines. Les éléments de cette image ont été générés grâce au logiciel Cytoscape [Shannon *et al.*, 2003]. On note la présence de deux composantes indépendantes et fortement connectées représentant le ribosome et l'ARN polymérase.

2.3.3 Conservation à travers les espèces

Certaines interactions ont été prédites séparément par plusieurs transferts en utilisant la méthode *InteroBH*. Elles proviennent donc de différentes espèces sources, ce qui nous fait penser qu'elles sont plus pertinentes [Lehner et Fraser, 2004]. De la même manière, plusieurs interactions sources ont été utilisées pour construire un unique lien entre des clusters de protéines orthologues au cours de la méthode *InteroPorc*. Ceci conduit à l'inférence d'interactions protéine-protéine prédites par plusieurs espèces, ce qui augmente la confiance que nous avons dans ces prédictions. C'est pourquoi nous avons voulu extraire ces interologues conservés à travers les espèces.

Les prédictions contiennent environ 5% d'interologues conservés au cours de l'évolution (voir Figure 2.5). Nous remarquons que le réseau contenant le plus d'interologues conservés, relativement à sa taille, est *InteroBH_HIGH*. Ceci est cohérent avec le fait que ce réseau a été défini en utilisant un seuil plus strict sur les comparaisons de séquences, sélectionnant ainsi les protéines puis les interologues les plus conservés. Parmi ces interactions, sept sont largement conservées puisqu'elles ont été transférées indépendamment par trois ou quatre espèces différentes.

Toutes les 258 interactions conservées sont représentées sur la figure 2.8 où on remarque deux protéines chaperone largement connectées. Ceci souligne le fait que les interactions avec des chaperones sont détectées par certaines méthodes expérimentales. Ce grand nombre d'interactions peut être dû à la fonction même des protéines chaperones qui est de se lier à un grand nombre de protéines pour assister leur repliement. Mais il peut être du à des interactions non spécifiques également. Notons que 75% des partenaires interagissant avec la protéine *groL1*(*slr2076*) ne partagent aucun terme GO avec elle. De plus, ces interactions ont été largement transférées à partir d'interactions identifiées par des méthodes à haut-débit, ce qui pourrait faire douter de leur pertinence. D'un autre côté, l'autre protéine chaperone fortement connectée *dnaK2* (*sll0170*) a des termes GO en commun avec 40% des ses partenaires. De plus, nous avons remarqué que ces interactions ont au contraire été transférées à partir d'interactions sources identifiées par différentes techniques expérimentales telles que la cristallographie à rayon X, le filtrage moléculaire (*molecular sieving*), ou encore des méthodes basées sur l'immunoabsorbance (*enzyme linked immunoabsorbent assay*). Ceci confère plus de pertinence à ces interactions.

2.3.4 Identification par différentes techniques expérimentales

alors intéressés aux techniques expérimentales utilisées pour mettre en évidence les interactions protéine-protéine que nous utilisons comme source de données. Ces méthodes d'identification sont décrites grâce à des termes de vocabulaire contrôlé défini par PSI [Kerrien *et al.*, 2007b] dans un schéma global rendant compte des interactions moléculaires. Ces termes sont définis dans un fichier au format OBO, un format standard de description des ontologies[4]. Il est alors possible de visualiser leur organisation grâce

[4]OBO : Open Biomedical Ontologie http ://www.obofoundry.org/

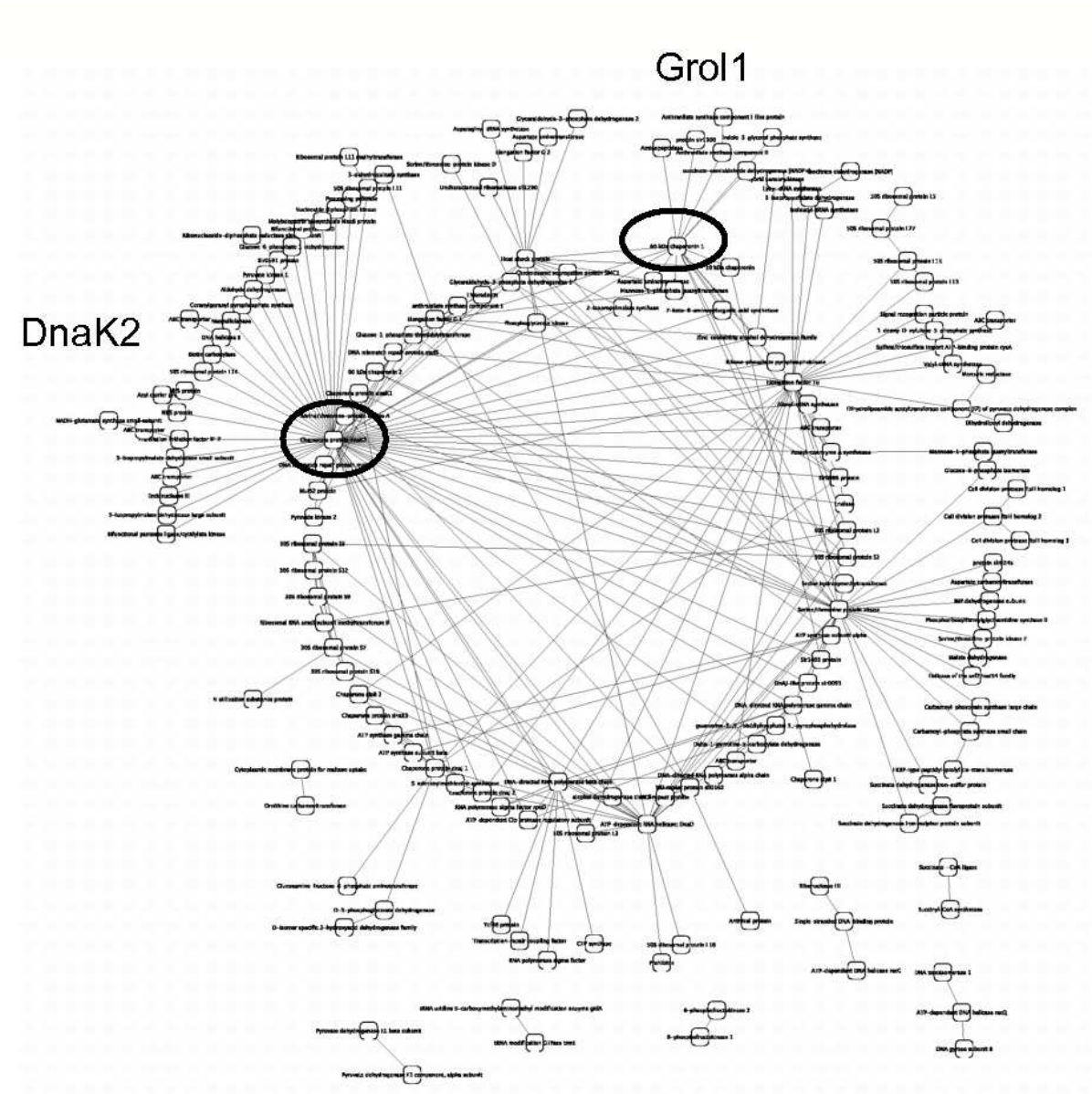


FIG. 2.8 – **Interactions conservées entre plusieurs espèces.** Ce réseau représente les 258 interactions (entre 167 protéines) qui ont été prédites séparément par plusieurs espèces. Ce graphe a été généré grâce au logiciel Cytoscape [Shannon *et al.*, 2003].

à l'outil Ontologie Lookup service (OLS)[5]. Nous avons voulu déterminer pour chaque interaction source si elle avait été identifiée par des méthodes différentes, voire par des approches expérimentales largement différentes.

En effet, chaque approche expérimentale a ses avantages et ses inconvénients [Hakes *et al.*, 2008] [von Mering *et al.*, 2002]. C'est pourquoi il est intéressant de savoir qu'une interaction a été identifiée par des approches différentes ; cela peut lui conférer une plus grande pertinence. Pour cela, nous avons défini des grands groupes de méthodes en regroupant tous les enfants, c'est-à-dire tous les termes plus spécifiques (voir Terminologie page 19) des termes suivants :

- MI :0401 : biochemical
- MI :0090 : Y2H
- MI :0013 : biophysical
- MI :0428 : imaging
- MI :0254 : genetic
- MI :0255 : transcription
- MI :0362 : inference
- MI :0686 : unspecified

Certains termes n'apparaissent pas dans les interactions sources tels que *unspecified*, *genetic* ou *transcription*.

Ainsi, nous avons calculé le nombre d'approches utilisées pour identifier chacune des interactions sources. 491 interactions ont été transférées à partir d'interactions sources identifiées par différentes techniques expérimentales. Le réseau *InteroBH_LOW* possède 6% d'interactions provenant de plusieurs techniques d'identification, alors que le réseau *InteroPorc* n'en contient que 4%.

2.3.5 Mise en évidence expérimentale

Parmi les interactions prédites, 10 étaient déjà décrites dans les interactions des bases de données IntAct, MINT et DIP qui contenaient en tout 185 interactions chez *Synechocystis* provenant d'expérience à bas-débit. Pour évaluer la pertinence de ce recouvrement, nous avons calculé la probabilité de trouver par hasard un recouvrement au moins aussi grand que celui observé. Nous avons utilisé pour cela un modèle hypergéométrique selon lequel la probabilité était inférieure à 10^{-4} (voir Annexe C.5). Par conséquent, les données expérimentales corroborent les prédictions.

De plus, une étude expérimentale à haut-débit a été conduite récemment chez *Synechocystis* [Sato *et al.*, 2007], mettant en évidence 3 236 interactions entre 1 920 protéines (voir Chapitre 3 pour une étude plus détaillée). Si on ne considère que les protéines qui font partie de cet ensemble, il reste 3 904 interactions prédites au lieu de 8 783. Parmi cet ensemble de 3 904 interactions, Sato *et al.* en ont identifiées 25, ce qui est également significatif (p-value $< 10^{-8}$). Il est important de noter que les jeux de données issus d'identification à haut-débit d'interactions protéine-protéine ne se recoupent que très peu jusqu'à présent. En effet, moins de 10% du nombre total d'interactions est

[5]OLS http ://www.ebi.ac.uk/ontology-lookup/browse.do ?ontName=MI

en commun entre deux études réalisées sur la même espèce et avec la même technique [Arifuzzaman *et al.*, 2006]. Ceci souligne le fort taux de faux négatifs de ces techniques, ce que nous retrouvons également ici lors de la comparaison avec les prédictions, et qui a été à nouveau souligné récemment [Yu *et al.*, 2008].

Au final, 35 interactions parmi les prédictions ont été identifiées expérimentalement et décrites dans les bases de données ou dans l'étude de Sato *et al.*Ces interactions sont représentées sur la figure 2.9.

2.3.6 Comparaison avec STRING

La base de données STRING met à disposition des prédictions de relations fonctionnelles entre protéines [von Mering *et al.*, 2003], [von Mering *et al.*, 2005], [von Mering *et al.*, 2007], [Jensen *et al.*, 2008]. Ces relations ne sont pas nécessairement des interactions physiques. En effet, elles proviennent de différentes méthodes de prédiction comme la co-expression, la co-localisation, la co-citation dans la littérature et l'identification expérimentale. De plus, toutes les relations fonctionnelles prédites sont également transférées, selon le principe des interologues, d'une espèce vers une autre. Ainsi, le transfert des relations identifiées expérimentalement correspond au principe de prédiction que nous avons utilisé.

Cette base de données met à disposition une interface web interactive qui permet d'explorer les interactions prédites pour certaines protéines. Néanmoins, il est difficile d'obtenir des résultats complets à l'échelle d'une espèce comme l'outil *InteroPorc* le permet. Certes, la base de données peut être obtenue dans le cas d'une utilisation académique dans une version antérieure à celle présentée sur l'interface web, mais il s'agit d'une utilisation plus complexe que l'outil *InteroPorc* en ligne de commande.

De plus, il est important de noter que la qualité des interactions protéine-protéine prédites dépend largement de la qualité des interactions sources. En effet, une interaction peut être détectée expérimentalement alors qu'elle ne se produit pas dans la réalité (faux positif). Si les deux protéines en interaction ont des orthologues dans l'espèce cible, cette interactions pourra être transférée. Ainsi, il est primordial de choisir précisément les interactions sources sur lesquelles va se baser le transfert. L'outil *InteroPorc* possède à ce sujet l'intérêt majeur d'implémenter la méthode indépendamment des données d'interactions sources. Il est donc possible de réaliser des transferts pour différents jeux de données sources. Ceci n'est pas possible avec la base de données STRING.

En définitive, la base de données STRING et l'outil de prédiction *InteroPorc* sont faits pour des utilisations et des objectifs différents. L'outil *InteroPorc* est assurément plus restreint puisqu'il ne prédit que des interactions physiques par la méthodes des interologues, mais il présente une utilisation beaucoup plus flexible et plus aisée pour cet objectif spécifique. En particulier, les prédictions sont possible avec *InteroPorc* pour toutes les espèces contenues dans la base de données Integr8 [Kersey *et al.*, 2005], soit plus de 750 en octobre 2008.

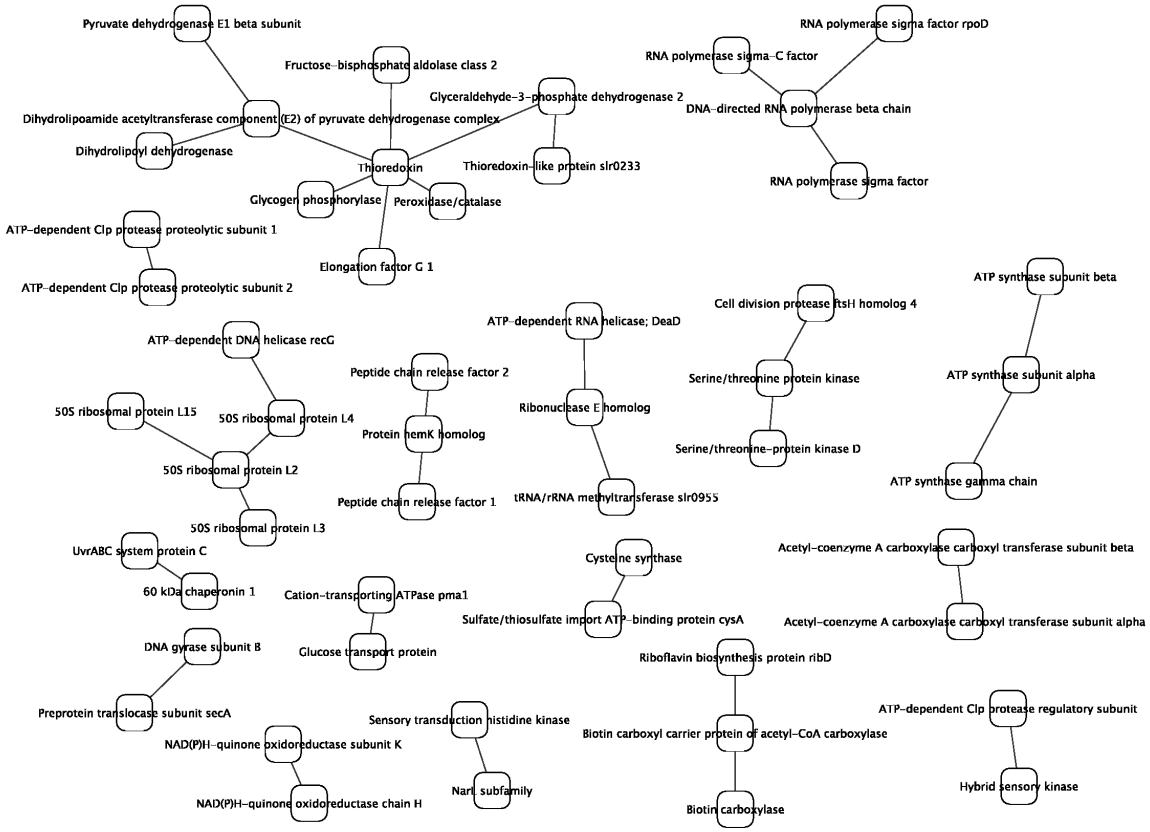


FIG. 2.9 – **Interactions confirmées expérimentalement.** Ce réseau représente les 35 interactions (entre 52 protéines) qui ont été mises en évidence expérimentalement. Ce graphe a été généré grâce au logiciel Cytoscape [Shannon *et al.*, 2003].

Parmi les 8 783 interactions protéine-protéine prédites chez *Synechocystis* et constituant le réseau *InteroFull*, nous avons extrait 3 496 interactions soutenues par différents types d'information comme les interactions domaine-domaine, les annotations fonctionnelles, la conservation des interologues à travers les espèces, la multiplicité des méthodes expérimentales ayant permis d'identifier les interactions sources, ou encore l'identification expérimentale des interactions.
Au cours de ce travail de prédiction *in-silico* d'interactions protéine-protéine, un outil a été développé et rendu disponible par une interface web.

2.4 Développement d'un outil de prédiction automatique d'interactions protéine-protéine

Un outil automatique de prédiction d'interactions protéine-protéine a été conçu, implémenté en langage java, puis mis à la disposition de la communauté lors de ce travail de thèse notamment grâce à une interface web. Nous faisons ici sa présentation et montrons quelques exemples d'application.

2.4.1 Introduction

Nous avons déjà vu que les interactions protéine-protéine sont d'un grand intérêt pour l'étude des processus biologiques. Mais les études expérimentales à grande échelle restent très coûteuses, entre autres en temps et en argent. Ainsi, seuls quelques organismes modèles ont pour l'instant des collections de tailles significatives d'interactions protéine-protéine identifiées expérimentalement. Le concept d'interologue est alors très utile pour transférer ces interactions vers d'autres organismes. Comme il a été décrit précédemment (voir page 54), de tels transferts ont déjà été faits pour un petit nombre d'organismes, mais chaque mise en application de ce concept reste manuelle, et les résultats disponibles sont alors des bases de données d'interactions transférées pour certains organismes. Les inconvénients sont, d'une part, qu'un organisme donné n'est pas nécessairement présent dans telle ou telle étude, et, d'autre part, que les données utilisées pour ces transferts ne sont plus modifiables. Même si certains résultats sont disponibles, les transferts sont effectués puis figés.

Il serait alors intéressant d'avoir un outil général plutôt que des bases de données. En effet, une telle application serait d'une grande utilité pour construire les premières cartes d'interactions pour les espèces qui n'en possèdent pas, ou encore pour compléter toute collection d'interactions plus ou moins importante. De plus, une méthode commune pour un grand nombre d'organismes pourrait faciliter grandement les études comparatives qui visent à mieux comprendre l'évolution des réseaux d'interactions à travers les espèces. Enfin, la notion d'outil est de grande importance car on peut choisir les données sources que l'on veut considérer. En effet, les interactions prédites dépendent fortement des interactions sources qui sont utilisées et l'on sait qu'il est difficile de quantifier la qualité des interactions actuellement disponibles.

Nous avons donc décidé, au cours de ce travail, de développer un tel outil, d'abord

dans l'intérêt de notre projet sur la cyanobactérie *Synechocystis* et, ensuite, pour toutes ces raisons précédemment exposées qui rendent une telle application très utile pour toute la communauté scientifique.

2.4.2 Présentation de l'outil automatique *InteroPorc*

L'outil *InteroPorc* est accessible grâce à une interface web où il peut être utilisé en ligne ou téléchargé pour une utilisation en ligne de commande. Cette interface web a été développée et est maintenue par Arnaud Martel, du Groupe Informatique pour les Scientifiques d'Ile de France (GIPSI) au CEA.

2.4.2.1 Accessibilité

Cet outil est un logiciel libre. Il est accessible de manière générale par l'intermédiaire d'une interface web[6] sous une licence Apache. Il peut être utilisé de différentes manières :

- Directement en ligne grâce à l'interface web
- En ligne de commande grâce à une librairie java contenant toutes les dépendances nécessaires
- En l'intégrant dans un programme java avec la librairie minimale sans les dépendances
- En utilisant le code source

Ce programme ayant été développé en langage java, il est utilisable sur toutes les plateformes.

2.4.2.2 Utilisation de l'interface web

L'utilisation la plus directe, la plus rapide, et sans doute la plus simple, est basée sur l'interface web. Les bases de données sources sont mises à jour régulièrement. Les versions utilisées sont précisées sur la page d'accueil (voir Figure 2.10). Certains résultats sont précalculés pour quelques espèces souvent utilisées comme *S. cerevisiae*, *E. coli*, *H. sapiens*, *A. thaliana*, *C. elegans*, *D. melanogaster*, *M. musculus*, *R. norvegicus* ou *H. pylori*. Les résultats sont alors consultables en ligne et téléchargeables.

Autrement, il suffit de donner l'identifiant de l'espèce souhaitée, et de lancer les prédictions en cliquant sur le bouton *run*. Cet identifiant est celui de la base de données NCBI Taxonomy [Wheeler *et al.*, 2008]. Les espèces disponibles sont répertoriées sur le site d'Integr8[7] où on pourra trouver cet identifiant (par exemple 1 148 pour *Synechocystis* 4 932 pour la levure ou 9 606 pour l'homme). Les prédictions sont faites en général en deux ou trois minutes.

Les résultats sont alors disponibles sur une page spécifique qui propose différents fichiers (voir Figure 2.11) consultables en ligne ou téléchargeables :

- *Predicted interactions* : il s'agit ici de la liste des interactions protéine-protéine prédites par *InteroPorc* pour l'espèce considérée. Ce fichier est disponible aux

[6] InteroPorc http ://biodev.extra.cea.fr/interoporc
[7] Integr8 http ://www.ebi.ac.uk/integr8

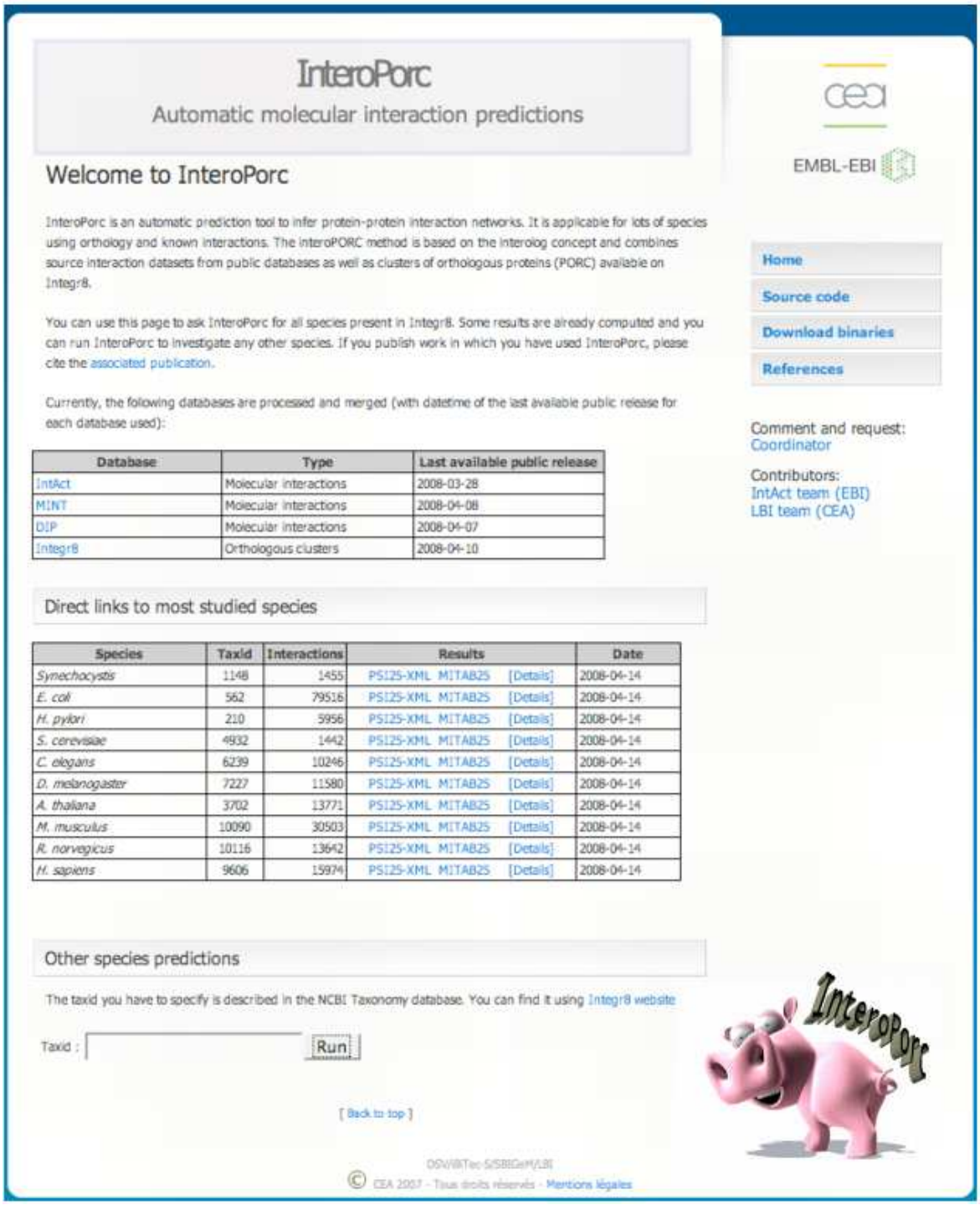


FIG. 2.10 – **Page principale de l'outil InteroPorc.** Cette image est une copie d'écran de la page principale de l'interface web disponible pour utiliser l'outil *InteroPorc* à l'adresse suivante http ://biodev.extra.cea.fr/interoporc.

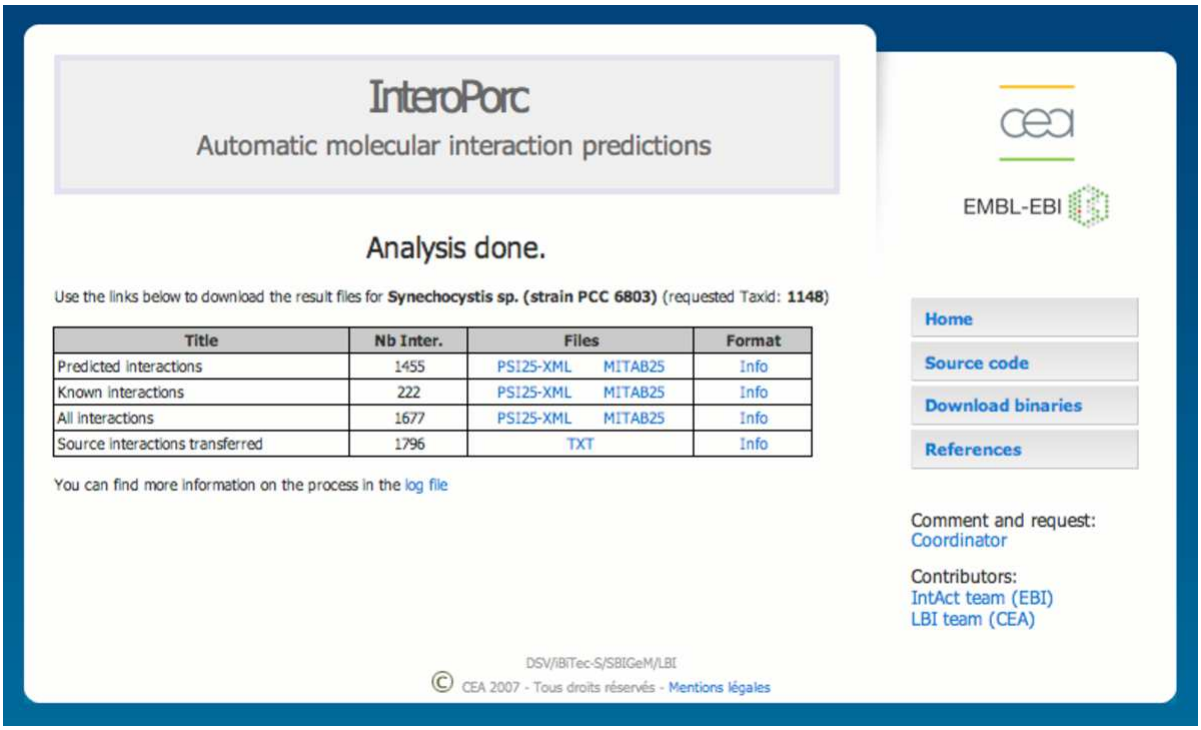


FIG. 2.11 – **Page de résultats de l'outil InteroPorc.** Cette image est une copie d'écran d'une page de résultats de l'interface web après utilisation de l'outil *InteroPorc* à l'adresse suivante http ://biodev.extra.cea.fr/interoporc. On trouve ici des fichiers contenant les interactions prédites et les interactions connues. Les interactions sont fournies dans deux formats standards définis par PSI : un format tabulé MITAB25 et un format XML PSI25-XML. De plus, un fichier récapitule quelques informations sur les interactions source utilisées lors du transfert. Sur la partie droite, on trouve des liens notamment pour télécharger l'outil, afin de l'utiliser en ligne de commande, ou pour consulter ou télécharger le code source. Il est aussi possible de consulter la publication associée [Michaut *et al.*, 2008d] et de contacter les auteurs par messagerie électronique.

formats standards MITAB[8] et PSI25-XML[9] définis par le consortium PSI [10] [Hermjakob *et al.*, 2004a].

- *Known & Predicted interactions* : nous ajoutons dans ce fichier les interactions connues pour l'espèce considérée provenant du dataset d'interactions sources. Les deux formats MITAB et PSI25-XML sont aussi disponibles.
- *Source interactions transferred* : dans ce fichier sont disponibles les interactions sources utilisées lors de chaque transfert. Il s'agit d'un simple fichier texte-tabulé.

2.4.2.3 Utilisation de l'application en ligne de commande

Pour utiliser l'application de manière indépendante, il faut la télécharger en cliquant sur le lien *Download binaries*, dans la partie droite de l'interface web (voir Figure 2.10). On obtient alors un fichier compressé (*interoporc.tar.gz*), qu'il faut donc décompresser et qui contient :

- *interoporc.jar* : la librairie java,
- *readme.txt* : un fichier d'informations (reproduit en Annexe F),
- *user.interoporc.log4j.properties* : un fichier de configuration,
- *LICENSE* : le fichier de license Apache.

Différentes options sont disponibles :

```
usage : Interoporc [OPTIONS]
Options :
 -o   –output-directory <file>      Directory where all files will be created
 -i   –mitab-file <file>            MITAB File, source interactions
 -p   –porc-file <file>             PORC file, orthologous clusters
 -x   –xml-files                    If output XML files are required
 -m   –max-nb-inter-xml <int>       Max nb of interactions to generate a XML
 -h   –help                         print this message
 -l   –lof-file <file>              use given file for log
 -t   –taxid <int>                  NCBI taxonomy identifier of the species
```

Pour faire des prédictions pour une espèce donnée, le plus simple est de suivre les étapes suivantes :

1. Créer un répertoire pour les prédictions (*dir*)
2. Déposer dans ce répertoire la librairie java (*interoporc.jar*)
3. Déposer dans ce répertoire un fichier au format MITAB avec toutes les interactions sources (*sourceInteractions.mitab*)
4. Déposer également dans ce répertoire le fichier de configuration pour les messages de sortie (*user.interoporc.log4j.properties*)

[8] Le format MITAB est un format texte tabulé.
[9] Le format PSI25-XML est un format XML.
[10] PSI : Proteomics Standard Initiative http ://www.psidev.info/

5. Choisir l'identifiant de l'espèce voulue (*taxid*)
6. Exécuter dans ce répertoire la commande suivante (voir d'autres exemples dans l'annexe F)

```
java -ms500m -mx1200m -cp interoporc.jar
uk.ac.ebi.intact.interolog.prediction.RunForOneSpecies -t 1 148
-i sourceInteractions.mitab -l user.interoporc.log4j.properties
```

Notons que les clusters de protéines orthologues sont téléchargés automatiquement s'ils ne sont pas présent dans le répertoire courant. Le fichier peut également être téléchargé sur le ftp d'Integr8[11] et donné en argument avec l'option -p. Une fois que l'application est ainsi lancée, des messages d'information s'affichent au fur et à mesure, et les fichiers de résultats sont créés dans ce répertoire.

2.4.3 Applications de l'outil *InteroPorc*

Dans un premier temps, nous avons appliqué l'outil *InteroPorc* à plusieurs espèces dans différents règnes. Dans un second temps, nous avons comparé les prédictions réalisées pour certaines espèces, avec les interactions mises en évidence expérimentalement.

2.4.3.1 Prédictions pour des espèces variées

Dans le cadre de notre projet, nous avons principalement utilisé l'outil *InteroPorc* pour construire un réseau d'interactions protéine-protéine pour la cyanobactérie *Synechocystis*. Néanmoins, ce dernier a une étendue bien plus importante puisqu'il peut s'appliquer à toutes les espèces ayant un génome séquencé et référencées dans la base de données Integr8 [Kersey *et al.*, 2005]. Nous présentons ici quelques exemples d'applications pour différents organismes représentatifs de la biodiversité des êtres vivants (voir Table 2.5).

Nous avons noté que le nombre d'interactions protéine-protéine prédites était plus élevé chez les eucaryotes que chez les bactéries et les archées. Ceci peut s'expliquer par la taille du protéome, qui est en général plus grande chez les eucaryotes, mais aussi par la distance, en termes d'évolution, entre les espèces sources et les espèces cibles. En effet, 83% des interactions sources ont été identifiées chez des eucaryotes. Or, il a été montré que le nombre d'interactions prédites par le concept d'interologue décroît quand la distance évolutive grandit entre les espèces source et cible [Brown et Jurisica, 2007]. Ainsi, il est plus aisé de transférer les interactions protéine-protéine dont nous disposons (majoritairement chez des eucaryotes) sur des eucaryotes plutôt que sur des procaryotes ou des archées.

La cyanobactérie *Anabaena* PCC7120 est un organisme modèle largement utilisé, dont le génome est séquencé, mais pour laquelle la carte d'interactions reste encore à

[11] ftp ://ftp.ebi.ac.uk/pub/databases/integr8/porc/proc_gene.dat

construire. En utilisant l'outil *InteroPorc*, nous pouvons prédire un ensemble de 1 678 interactions protéine-protéine, ce qui est un apport réellement intéressant (voir Table 2.5).

2.4.3.2 Comparaison avec les interactions connues

Dans la mesure où l'outil *InteroPorc* peut s'appliquer à un grand nombre d'espèces, il était naturel de faire des prédictions pour certaines espèces pour lesquelles un grand nombre d'interactions a été mis en évidence expérimentalement. Par définition, l'outil *InteroPorc* ne peut pas prédire des interactions qui sont présentes dans le jeu de données source. Ainsi, pour chaque espèce, nous avons réalisé les prédictions en retirant toutes les interactions connues du jeu de données source pour l'espèce en question. Les résultats sont indiquées dans la Table 2.6.

Les recouvrements entre les prédictions et les interactions mises en évidence expérimentalement sont faibles. Ceci provient sans doute de plusieurs choses. D'abord, les méthodes de prédictions ne sont par nature pas exhaustives, puisqu'elle se limitent au transferts d'interactions connues dans d'autres espèces. De plus, nous avons seulement considéré des homologies de séquences fortes pour limiter les faux positifs. Ceci réduit aussi le nombre d'interactions prédites. Enfin, il faut noter que les jeux de données expérimentaux sont également très incomplets [Yu *et al.*, 2008]. Même entre deux jeux de données expérimentaux utilisant la même technique, les recouvrements sont inférieurs à 10% pour l'instant [Arifuzzaman *et al.*, 2006].

Nous avons développé un outil automatique, *InteroPorc*, qui permet d'obtenir rapidement un réseau d'interactions protéine-protéine. Ceci est utile en particulier pour toute espèce récemment séquencée ou n'ayant pas ou peu d'interactions identifiées jusqu'à présent, ou encore pour compléter toute collection d'interactions déjà disponible. Cela facilitera notre compréhension des interactions protéine-protéine chez de nombreuses espèces, et pourra donner des pistes de validations expérimentales.

Règne	**Espèce**	**Identifiant de l'espèce**	**Protéome**	**Interactions**	
				décrites	**prédites**
Archée	*P. kodakaraensis*	69 014	2 301	0	221
Archée	*T. volcanium*	273 116	1 523	0	208
Eucaryote	*R.norvegicus*	10 116	12 028	2 178	13 469
Eucaryote	*A. fumigatus*	330 879	9 629	0	17 225
Eucaryote	*P. falciparum*	36 329	5 283	2 737	4 026
Bactérie	*B. subtilis*	224 308	4 105	0	2 160
Bactérie	*Synechocystis sp.*	1 148	3 506	185	1 463
Bactérie	*Anabaena sp.*	103 690	6 070	1	1 678

TAB. 2.5 – **Interactions nouvelles prédites par *InteroPorc*.** Pour chaque espèce, les colonnes indiquent le règne suivi du nom de l'espèce. La colonne suivante donne l'identifiant taxonomique de cette espèce fourni par la base de données NCBI Taxonomy. Sont ensuite indiqués la taille du protéome de cette espèce (le nombre de protéines), le nombre d'interactions décrites dans le jeu de données source créé à partir des trois bases de données IntAct, MINT et DIP, puis le nombre d'interactions prédites par l'outil *InteroPorc*. Par définition, ces ensembles d'interactions sont disjoints dans la mesure où l'outil *InteroPorc* ne peut pas prédire des interactions présentes dans le jeu de données source.

Espèce	**Expérimental**	**Prédictions**	**Intersection**
Synechocystis	185	1 476	13
Saccharomyces cerevisiae	56 007	1 767	308
Caenorhabditis elegans	6 466	11 014	75
Drosophilia melanogaster	50 815	13 570	123
Homo sapiens	32 321	17 835	570
Arabidopsis thaliana	1 694	13 756	14
Helicobacter pylori	1 552	5 169	1
Escherichia coli	22 791	1 486	12

TAB. 2.6 – **Interactions connues prédites par *InteroPorc*.** Pour chaque espèce, les colonnes indiquent le nombre d'interactions présentes dans le jeu de données source créé à partir des trois bases de données IntAct, MINT et DIP (colonne **Expérimental**), puis le nombre d'interactions prédites par l'outil *InteroPorc* (colonne **Prédictions**), et enfin le nombre d'interactions en commun dans ces deux ensembles (colonne **Intersection**). Pour chaque prédiction, les interactions de l'espèce en question ont été retiré du jeu de données source afin de permettre leur éventuelle prédiction.

Conclusion

Dans ce chapitre, nous avons développé des méthodes de prédiction *in-silico* d'interactions protéine-protéine basées sur la conservation entre les organismes. Ainsi, le principe des interologues permet de transférer la connaissance des interactions chez un ou plusieurs organismes vers un ou plusieurs autres organismes. Ce principe général peut s'appliquer à toutes les espèces. C'est pourquoi nous avons développé un outil automatique de prédiction qui permet d'obtenir un réseau d'interactions protéine-protéine pour tous les organismes dont le génome est séquencé et présent dans la base de données Integr8. Néanmoins, le transfert automatique ne s'effectue que pour les protéines très conservées, produisant par conséquent des réseaux incomplets.

Toutefois, dans le cas de la cyanobactérie *Synechocystis*, nous avons développé une méthode complémentaire. Cette méthode est plus flexible car elle permet de définir différents niveaux de confiance sur les transferts effectués. Ainsi, la combinaison de ces approches nous a permis de construire un réseau d'interactions protéine-protéine chez *Synechocystis* appelé *InteroFull*. De plus, nous avons extrait un ensemble d'interactions protéine-protéine soutenues par d'autres données, comme par exemple les interactions domaine-domaine ou les annotations fonctionnelles. Cependant, l'évaluation des interactions prédites reste qualitative et il serait intéressant de réaliser une évaluation quantitative de chacune des interactions prédites.

Par ailleurs, il est important de noter que la qualité des interactions prédites dépend largement de la qualité des interactions sources. Or, les jeux de données d'interactions protéine-protéine identifiées expérimentalement sont actuellement de qualités variables, avec un grand nombre de faux positifs. Ainsi, il est primordial de sélectionner les interactions protéine-protéine sur lesquelles se base le transfert. C'est pourquoi nous avons implémenté la méthode *InteroPorc* de manière indépendante des données. Il est alors possible d'appliquer l'outil à des jeux de données choisis.

Enfin, l'outil *InteroPorc* que nous avons développé est disponible librement sur internet. Il peut être téléchargé ou utilisé en ligne. Il est d'un grand intérêt pour avoir une première image des réseaux d'interactions protéine-protéine de toutes les espèces séquencées.

Ce travail a donné lieu à trois publications dont une dans un journal scientifique [Michaut *et al.*, 2008d] (voir l'article à la fin du manuscrit) et deux autres dans des actes de colloques [Michaut *et al.*, 2007], [Michaut *et al.*, 2008b]. Un poster a également été sélectionné lors d'un congrès international [Michaut *et al.*, 2008c] (voir la liste des publications page 260).

En même temps que ce travail de prédiction était effectué, des données expérimentales ont été publiées pour *Synechocystis*. Par conséquent, l'idée suivante a été d'analyser ces données expérimentales et de les comparer avec les prédictions.

Chapitre 3

Comparaison des interactions prédites avec les données expérimentales

"La biologie occupe, parmi les sciences, une place à la fois marginale et centrale."

Jacques Monod,
Le hasard et la nécessité, 1973

Dans ce chapitre, notre objectif était de comparer les interactions protéine-protéine provenant des méthodes de prédiction *in-silico* que nous avons développées,et des observations expérimentales récemment publiées [Sato *et al.*, 2007]. Nous voulions notamment considérer les organisations globales et locales des réseaux d'interactions protéine-protéine. Or, avant de comparer différents jeux de données, il est conseillé de les étudier séparément afin d'en identifier les forces et les faiblesses [Aloy, 2007], [Gentleman et Huber, 2007]. C'est pourquoi, nous avons d'abord voulu analyser les données expérimentales pour *Synechocystis* en les comparant aux autres études expérimentales à grande échelle réalisées précédemment, en particulier chez la levure. Les interactions prédites ont quant à elles été analysées au cours du chapitre précédent. Nous avons donc ensuite comparé les interactions obtenues par les méthodes de prédiction et par l'identification expérimentale. Puis, nous avons analysé la structures des réseaux en nous intéressant à leur topologie. Enfin, nous avons extrait les modules fonctionnels des différents réseaux avec une méthode commune et comparé les résultats.

3.1 Analyse des données expérimentales

Avant de comparer les données expérimentales et prédites, nous avons voulu analyser les données expérimentales séparément, notamment en les comparant à d'autres jeux de données expérimentaux. Nous avons d'abord choisi un modèle de représentation des données. Nous avons ensuite sélectionné les principaux jeux de données disponibles dans la littérature permettant une comparaison pertinente et exhaustive. Nous avons alors voulu étudier la couverture de l'approche expérimentale menée par Sato *et al.* en termes de protéines et en termes d'interactions [Sato *et al.*, 2007]. Enfin nous avons étudié l'asymétrie de la technique d'identification présente dans les différentes études expérimentales, notamment en la quantifiant au niveau des interactions, puis au niveau des protéines.

3.1.1 Représentation des données

Dans le cas des interactions considérées jusqu'à présent, nous avons modélisé chaque interaction par un arc non orienté entre deux nœuds représentant chacun une protéine (voir page 52). Un réseau d'interactions protéine-protéine était alors représenté sous la forme d'un graphe non orienté, puisque la notion d'interaction *in-vivo* est symétrique.

Dans le cas des interactions mises en évidence expérimentalement, le concept d'interaction protéine-protéine reste symétrique. Néanmoins, les techniques expérimentales utilisées pour mettre en évidence ces interactions introduisent des rôles différents pour les protéines (appât, proie). Les interactions détectées ne sont donc pas symétriques. Ainsi, une interaction peut être observée de A vers B (si A est l'appât et B la proie) et non de B vers A (si B est l'appât et A la proie). Cette information nous renseigne sur les limitations de la technique utilisée, et elle est donc intéressante à considérer. C'est pourquoi nous avons choisi de représenter les données expérimentales par des graphes orientés. Ainsi, une interaction détectée entre l'appât A et la proie B est représentée par un arc orienté de A vers B.

De plus, nous avons utilisé les concepts d'appâts et de proies viables définis par Chiang *et al.* [Chiang *et al.*, 2007] : un appât viable est un appât (*Bait*) qui a détecté au moins une proie ; une proie viable est une protéine qui a été détectée en tant que proie (*Prey*) par au moins un appât. Ainsi, nous avons défini les ensembles VB (*Viable Bait*) et VP (*Viable Prey*) selon la définition 3.1.

Définition 3.1 (VB et VP) *Soient $\mathcal{P}$, le protéome de l'espèce considérée, $B \subset \mathcal{P}$ l'ensemble des appâts, $P \subset \mathcal{P}$ l'ensemble des proies et $G = (E, V)$, le graphe d'interactions, où $E = B \times P$ est l'ensemble des interactions orientées.*

(i) $VB = \{p \in B \mid \exists\, a \in P, (p, a) \in E\}$
(ii) $VP = \{p \in P \mid \exists\, b \in B, (b, p) \in E\}$

De plus, nous avons défini, à partir de ces ensembles, les trois ensembles VBP, VBO et VPO permettant d'extraire les protéines qui sont à la fois appâts et proies viables (voir Définition 3.2).

Définition 3.2 (VBO, VPO et VBP)
(i) $VBP = VB \bigcap VP$ (*Viable Bait and Prey*)
(ii) $VBO = VB \setminus VBP$ (*Viable Bait Only*)
(iii) $VPO = VP \setminus VBP$ (*Viable Prey Only*)

Après avoir établi cette représentation des réseaux d'interactions protéine-protéine, nous avons voulu sélectionner différents jeux de données pertinents, afin de les comparer avec les interactions mises en évidence par Sato *et al.* [Sato *et al.*, 2007].

3.1.2 Sélection et description des jeux de données

Sato *et al.* ont réalisé une identification à haut-débit des interactions protéine-protéine chez *Synechocystis* en utilisant une approche par double-hybride chez la levure [Sato *et al.*, 2007]. Ainsi, les auteurs ont mis en évidence 3 236 interactions entre 1 920 protéines (voir Figure 3.1). Les interactions observées ont été classées en quatre catégories (A à D) pour indiquer la fiabilité des données [Sato *et al.*, 2007]. Cette classification dépend du nombre de clones proies positifs soutenant l'interaction (voir page 48). Ainsi, les catégories A et B incluent les interactions soutenues par plusieurs clones. La catégorie C est quant à elle constituée des interactions soutenues par un seul clone. Enfin, la catégorie D regroupe les interactions soutenues par des clones proies ayant été détectés par au moins 18 appâts. Pour cette raison, les interactions de la catégorie D sont considérées comme douteuses.

À partir de ces données, nous avons défini deux sous-ensembles (voir Définition 3.3) : *SatoFull* contient toutes les interactions identifiées par Sato *et al.*, et *SatoCore* contient seulement les interactions des catégories A et B (voir Figure 3.2).

Définition 3.3 *Soient I, la liste des interactions identifiées par Sato* et al.*, et cat, la fonction $I \rightarrow \{A, B, C, D\}$ qui associe à chaque interaction sa catégorie,*

(i) $i \in \text{SatoFull} \iff cat(i) \in \{A, B, C, D\}$
(ii) $i \in \text{SatoCore} \iff cat(i) \in \{A, B\}$

Pour mener à bien notre analyse comparative, nous avons considéré un ensemble de jeux de données d'interactions protéine-protéine identifiées chez la levure et la bactérie *H. pylori* (voir Annexe G et Figure 3.3). Ces jeux de données ont été construits à partir des deux approches expérimentales suivantes : le double-hybride chez la levure (*Y2H*) et la spectrométrie de masse de complexes purifiés (*AP-MS*). Rappelons que pour une approche donnée, chaque mise en œuvre est différente (voir page 47).

Parmi ces jeux de données, certains possèdent aussi un sous-ensemble considéré de meilleure qualité et dénoté *CORE*. Le jeu de données décrit par Ito *et al.* contient 4 449 interactions entre 3 242 protéines de levure (*ItoFull*). Parmi ces interactions, celles qui ont été détectées par au moins trois IST (*Interaction Sequence Tag*) constituent le jeu de données *ItoCore*.

Le jeu de données produit pour la bactérie *H. pylori* était particulièrement intéressant dans le cadre de cette comparaison avec les données de Sato *et al.* dans la mesure où la

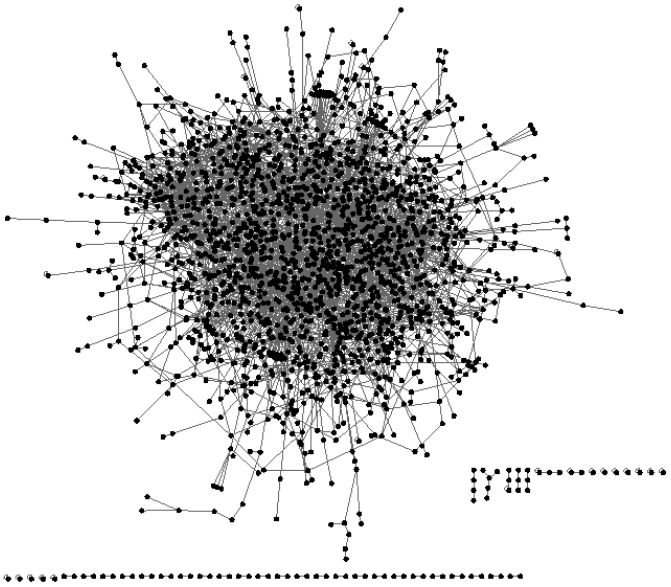


FIG. 3.1 – **Réprésentation graphique du jeu de données *SatoFull*.** Nous avons utilisé le logiciel Cytoscape [Shannon *et al.*, 2003] pour représenter les 3 236 interactions entre 1 920 protéines mises en évidence par Sato *et al.* Le graphe est ici non orienté car le détail des interactions n'est pas l'objet de ce graphique (et n'est pas visible).

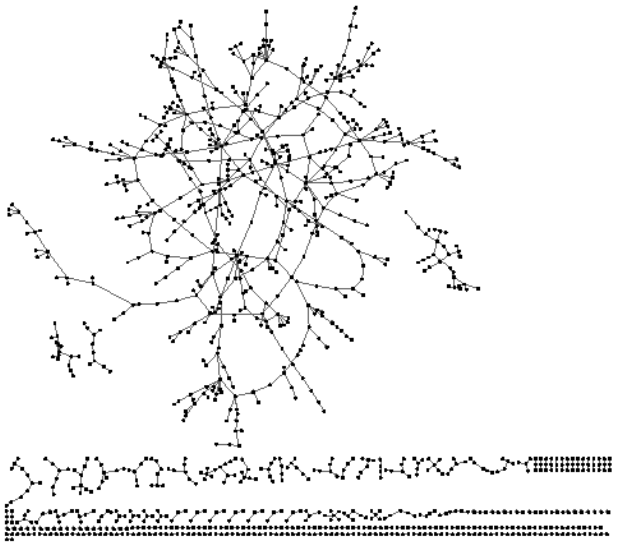


FIG. 3.2 – **Représentation graphique du jeu de données *SatoCore*.** Nous avons utilisé le logiciel Cytoscape [Shannon *et al.*, 2003] pour représenter les 1 064 interactions de catégories A et B entre 1 152 protéines mises en évidence par Sato *et al.* Le graphe est ici non orienté car le détail des interactions n'est pas l'objet de ce graphique (et n'est pas visible).

technique expérimentale et l'organisme d'étude sont proches de ceux utilisés par Sato *et al.* Le jeu de données *RainFull* est constitué de 1 568 interactions entre 740 protéines. Ces interactions sont classées en cinq catégories (A à E). Les catégories A à D sont définies en fonction du score dénoté PBS (*PIM biological score*) [Rain *et al.*, 2001]. Ce score est basé sur un modèle statistique de la compétition entre les fragments pour se lier à l'appât. La cinquième catégorie E a été ajoutée afin de distinguer les interactions impliquant des proies considérées comme douteuses. Cette démarche a été reprise par Sato *et al.* quand ils ont défini leur catégorie D. Nous avons alors défini le jeu de données *RainCore* comme l'ensemble des interactions de catégories A et B (voir Définition 3.4).

Définition 3.4 *Soient I, la liste des interactions identifiées par Rain* et al.*, et cat, la fonction $I \rightarrow \{A, B, C, D\}$ qui associe à chaque interaction sa catégorie,*

(i) $i \in \text{RainFull} \iff cat(i) \in \{A, B, C, D, E\}$
(ii) $i \in \text{RainCore} \iff cat(i) \in \{A, B\}$

Nous avons donc voulu comparer les interactions protéine-protéine mises en évidence par Sato *et al.* avec l'ensemble de ces jeux de données expérimentaux. L'objectif était tout d'abord d'évaluer la couverture de cette étude en termes de paires de protéines testées, et de la comparer avec les couvertures des autres études.

3.1.3 Évaluation de la couverture

Notre premier objectif était de comparer la couverture des différentes études, c'est-à-dire le nombre d'interactions testées par rapport à la taille du protéome. Néanmoins, cette information n'est en général pas accessible directement. En effet, l'identification expérimentale d'une interaction consiste à inférer son existence à partir des données mesurées, ces dernières dépendant de la technologie mise en œuvre. Cette inférence peut être vraie (vrai positif) ou fausse (faux positif). En général, une observation positive indique l'existence d'une interaction. Mais cette approche ne prend pas en compte une information essentielle qui est l'ensemble des interactions testées. En effet, la plupart des jeux de données disponibles publiquement indiquent les mesures positives, mais pas les mesures négatives. Dans ce cas, les paires non testées ne sont pas distinguables des paires négatives.

Or, quelle que soit la technologie utilisée, certaines limitations techniques font que certaines interactions ne peuvent pas être testées. Par exemple, dans le cas du *Y2H* où les protéines testées sont liées à des domaines, seules les constructions permettant d'obtenir une protéine de fusion fonctionnelle peuvent être testées. Il est également difficile d'identifier des interactions faisant intervenir des protéines membranaires dans la mesure où le test se déroule dans le noyau. Ainsi, même les études décrites comme ayant été réalisées à l'échelle du génome ne permettent pas de tester toutes les paires possibles de protéines. Néanmoins, le détail des paires de protéines réellement testées n'est en général pas connu ou pas disponible. C'est pourquoi nous avons utilisé les notions d'appâts et de proies viables définies dans la section 3.1.1, afin d'estimer, d'une part les

protéines qui ont pu servir d'appâts ou être détectée comme proie, et d'autre part les interactions qui ont été testées.

3.1.3.1 Couverture du protéome

En utilisant les concepts d'appâts et de proies viables définis dans la section 3.1.1, nous avons identifié dans *SatoFull* 1 044 proies viables (VB) et 1 352 appâts viables (VP), parmi lesquels 476 protéines sont à la fois proie et appât viable (VBP). Parmi les 3 650 protéines de *Synechocystis*, 52% sont inclues dans cette étude (15% comme appât viable seulement (VBO), 24% comme proie viable seulement (VPO) et 13% comme appât et proie viable (VBP)), et les 48% restant n'ont pas été testés (voir Figure 3.3).

Nous avons comparé cette répartition du protéome avec les autres jeux de données (voir Figure 3.3). Nous avons alors noté que l'étude de Sato *et al.* se situait parmi les meilleures études en termes de couverture relative du protéome. De plus, nous avons remarqué que les répartitions des protéines de *SatoFull* et *SatoCore* étaient très proches de celles de *RainFull* et *ItoFull*.

Par ailleurs, nous avons noté que le protéome n'était pas couvert de manière aléatoire dans l'étude de Sato *et al.*, puisque les protéines appâts ont été sélectionnées sur la base de différents critères. D'un côté, Sato *et al.* ont sélectionné les 1 832 appâts selon des critères biologiques. En effet, ils ont choisi les gènes faisant partie des groupes suivants :

- Gènes potentiellement impliqués dans la transduction du signal pour les mécanismes de contrôle de la phosphorylation (*two-component system*) (5%)
- Gènes dont les homologues sont conservés dans le génome d'*A. thaliana* (36%)
- Gènes de fonction inconnue (57%)

Aucun critère de choix n'a été spécifié pour les autres appâts (2%).

D'un autre côté, Sato *et al.* ont utilisé un critère technologique : ils ont retiré tous les gènes codant des protéines avec des domaines transmembranaires, à cause de la difficulté de détecter des interactions pour de telles protéines en utilisant l'approche double-hybride chez la levure. Par conséquent, le protéome n'est pas couvert aléatoirement mais sujet à un biais de sélection.

Cette analyse nous a permis de déterminer le rôle des différentes protéines, et en particulier celles étant viables à la fois comme appâts et comme proies (VBP). Nous avons alors pu évaluer la couverture des différentes études en termes d'interactions.

3.1.3.2 Couverture de l'interactome

Après avoir analysé la couverture des différentes études en termes de protéines, l'objectif était de comparer les couvertures en termes d'interactions testées. En théorie, les interactions testées sont celles étant constituées d'une protéine ayant servi d'appât et de n'importe quelle autre protéine, quand une approche par criblage de banque est utilisée. Dans la pratique, certains appâts ne fonctionnent pas, c'est-à-dire qu'ils ne détectent aucune proie pour des raisons techniques. D'autre part, certaines protéines peuvent ne pas être détectées comme proie, pour des raisons techniques également, comme par exemple les protéines membranaires. Par conséquent, nous avons évalué l'espace des interactions testées en considérant celles dont nous étions certains qu'elles avaient été testées. Pour

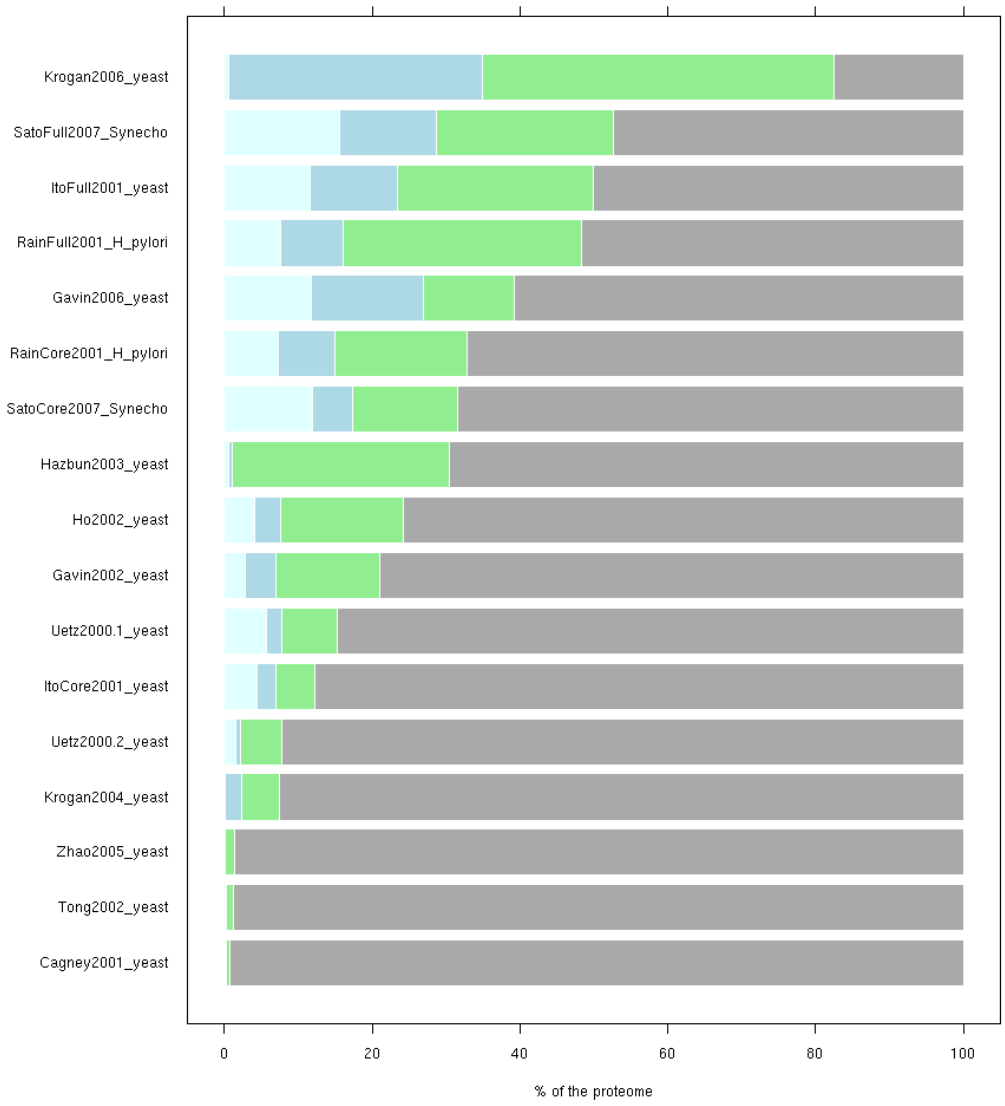


FIG. 3.3 – **Répartition des protéines pour un ensemble de jeux de données à grande échelle.** Cette figure montre la répartition des protéines testées et non testées dans les jeux de données considérés (et décrits dans l'Annexe G). Toutes les protéines du protéome de l'espèce considérée sont représentées ici (en pourcentage) : les appâts viables seulement (VBO) en bleu clair, les appâts/proies viables (VBP) en bleu foncé, les proies viables seulement (VPO) en vert, et les protéines non testées en gris.

cela, il fallait que la protéine appât soit viable (VB), et que la protéine proie soit viable également (VP).

Notons que chaque *paire orientée* de protéines peut être :

- Testée sans interaction détectée
- Testée avec une interaction détectée
- Non testée

Nous avons donc évalué la couverture de l'interactome en calculant le pourcentage des interactions testés par rapport à l'espace total des paires orientées de protéines (voir Définition 3.5) :

Définition 3.5 (Couverture de l'interactome) *Soit $\mathcal{P}$, le protéome de l'espèce considérée. La couverture de l'interactome est définie comme le pourcentage d'interactions testées, par rapport à l'ensemble des couples de protéines :*

$$C = 100 \times \frac{|VB \times VP|}{|\mathcal{P} \times \mathcal{P}|} \tag{3.1}$$

Il faut noter que cette estimation peut être fausse dans les deux sens. En effet, certaines interactions ont pu être testées entre des protéines n'ayant donné lieu à aucune interaction observée. Les protéines ne sont alors pas viables et l'interaction n'est pas considérée comme testée. D'autre part, le caractère viable d'une protéine dépend en réalité de différents paramètres. Les conditions expérimentales peuvent par exemple être adaptées à chaque protéine appât. Ainsi, une protéine p, qui a été détectée par l'appât a, dans les conditions expérimentales A, ne sera peut-être pas viable dans les conditions expérimentales B, utilisées lors du test de l'appât b. Par conséquent, l'interaction $b \rightarrow p$ n'a pas réellement été testée, mais elle sera comptée comme testée par l'estimateur.

Concernant l'étude menée par Sato *et al.* sur *Synechocystis*, nous avons conclu que 11% des interactions avaient été testées. Cette couverture est similaire à celle de *ItoFull* (voir Figure 3.4).

3.1.3.3 Comparaison des couvertures des jeux de données restreints

Le jeu de données restreint de l'étude de Sato *et al.*, dénommé *SatoCore*, a été représenté sur la Figure 3.2. Il contient 1 064 interactions entre 1 152 protéines (631 VB et 716 VP dont 195 VBP, voir la Table 3.1). Comme les deux figures 3.1 et 3.2 le suggèrent, ce graphe est beaucoup moins connecté que celui de *SatoFull*. En effet, les deux critères suivants ont permis de le montrer :

- Le degré moyen est plus faible (1,7 au lieu de 3,2)
- Le diamètre de la principale composante connexe est plus grand (26 au lieu de 16)

Nous avons d'abord noté que les rapports entre *ItoCore* et *ItoFull* sont assez différents des deux autres jeux de données. Ceci peut s'expliquer par le fait que *ItoFull* a été obtenu par une approche matricielle, alors que *SatoFull* et *RainFull* ont été obtenus par une approche de criblage de banque (voir page 51).

D'autre part, les rapports de tailles ente les graphes *SatoCore* et *SatoFull* sont très proches de ceux entre *RainCore* et *RainFull* en termes de nombre de protéines et d'interactions. En revanche, les répartitions des protéines sont fortement différentes : la

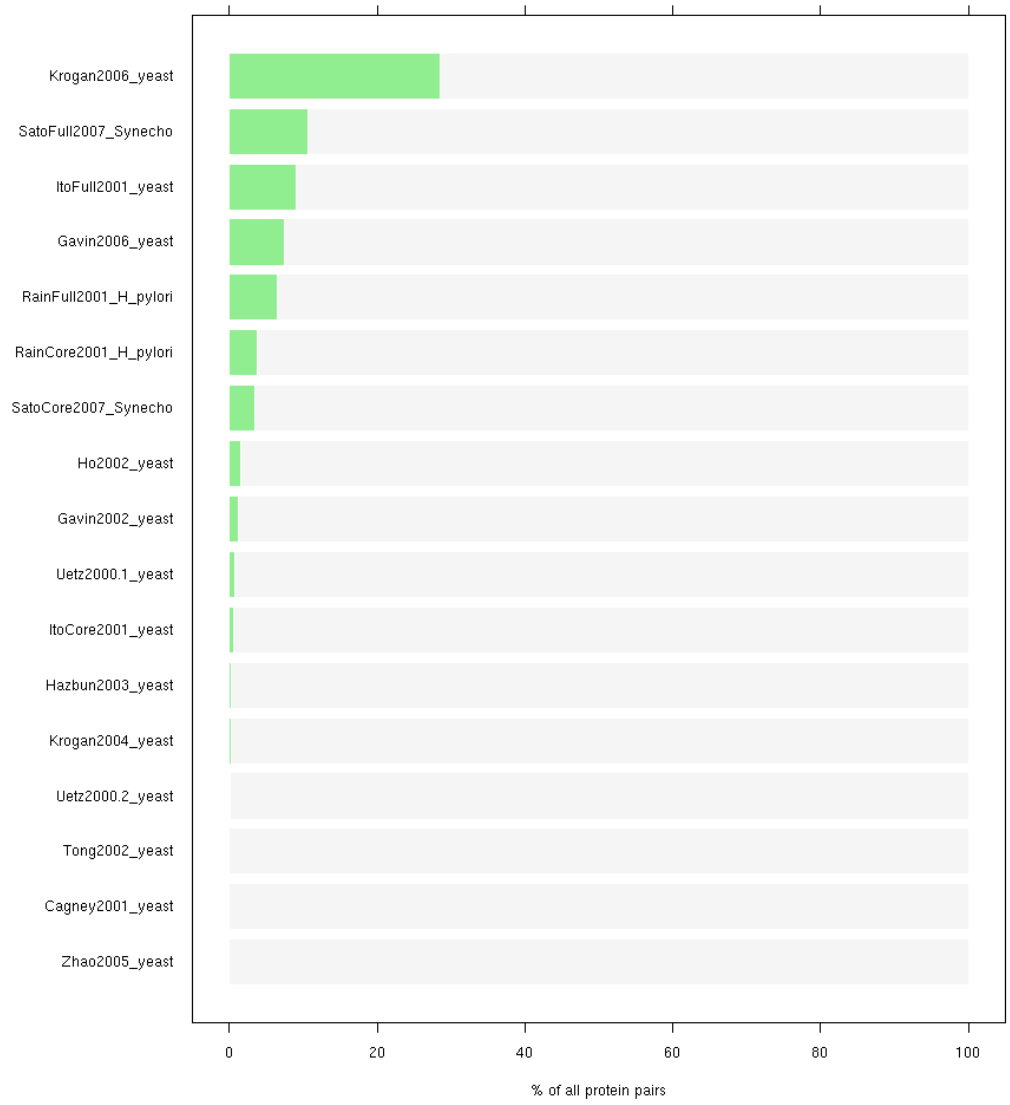


FIG. 3.4 – **Couverture de l'interactome pour un ensemble de jeux de données à grande échelle.** Cette figure montre le taux de couverture de l'interactome pour les différents jeux de données considérés (et décrits dans l'Annexe G). Il s'agit du pourcentage de paires orientées de protéines testées dans l'étude, par rapport à l'ensemble des paires orientées de protéines de l'espèce considérée.

restriction de *RainFull* à *RainCore* entraîne une forte diminution de la part de protéines seulement proies viables. Par contre, la restriction de *SatoFull* à *SatoCore* entraîne une diminution homogène des protéines dans chaque catégorie. Ces différences peuvent provenir de nombreuses subtilités techniques, comme par exemple le processus de sélection des appâts et leur qualité, ou les conditions de sélection des proies et des clones séquencés. Ceci montre l'importance de bien comprendre la finesse technique des approches expérimentales. Les informations détaillées sur le montage des cribles sont primordiales. Néanmoins, elles sont rarement détaillées comme cela a déjà été souligné par Huang *et al.* [Huang *et al.*, 2007a].

À cette étape, nous voulions étudier les biais systématiques contenus dans ces données expérimentales, et en particulier l'asymétrie provenant des méthodes de détection.

3.1.4 Analyse de l'asymétrie des méthodes de détection

Nous avons déjà souligné le fait que les études expérimentales, même à grande échelle, étaient soumises, dans la majorité des cas, à un biais de sélection. De plus, les technologies utilisées induisent également des biais. L'approche double-hybride, par exemple, peut être confrontée à des problèmes d'ordre technique, comme par exemple l'auto-activation du gène rapporteur pour certains appâts, ou encore les protéines dites collantes qui sont détectées comme pouvant interagir avec de nombreuses autres protéines, sans que ces interactions aient nécessairement lieu *in vivo*. Les rôles différents des appâts et des proies lors de la détection expérimentale d'une interaction, ainsi que les problèmes techniques évoqués, induisent une asymétrie dans la détection des interactions. Nous avons alors voulu quantifier cette asymétrie à deux niveaux différents : au niveau des interactions et au niveau des protéines.

3.1.4.1 Asymétrie au niveau des interactions

D'abord, l'idée a été de quantifier l'asymétrie de manière globale, sur un jeu de données complet, en considérant l'ensemble des interactions. Considérons une interaction donnée entre deux protéines A et B. Si la technologie de détection expérimentale était parfaite, on devrait, soit identifier les deux interactions orientées (A→B et B→A), soit n'en détecter aucune. L'asymétrie de la technique de détection s'exprime dans le fait qu'on trouve dans les données expérimentales des interactions orientées (par exemple A→B) sans trouver leur réciproque (B→A). Dans ce cas, on peut en conclure que l'interaction détectée A→B est une erreur de détection (par exemple problème d'auto-activation), ou bien que l'interaction B→A n'a pas pu être détectée à cause d'un problème technique, alors qu'elle aurait du être détectée (par exemple un problème de sous-échantillonnage des clones séquencés). Nous avons donc choisi de considérer cet aspect réciproque des interactions pour quantifier l'asymétrie des différents jeux de données.

Pour cela, nous avons tout d'abord restreint l'ensemble des interactions considérées. En effet, nous devions être certains que les deux interactions orientées (A→B et B→A) avaient bien été testées. C'est pourquoi nous avons restreint notre analyse aux interactions ayant lieu entre des protéines du sous-ensemble VBP. De plus, les interactions entre

deux instances d'une même protéine, représentées par une boucle (voir page 52), n'apportent aucune information utile quant à la symétrie de la technologie. Elles ne peuvent pas être considérées comme détectées de manière réciproque, car une seule interaction est réellement détectée, mais elles ne peuvent pas non plus être considérées comme des interactions non-réciproques. Par conséquent, nous les avons également filtrées.

Nous avons alors défini le taux de symétrie S comme le pourcentage d'interactions détectées de manière symétrique par rapport à l'ensemble des interactions entre des protéines VBP (voir Définition 3.6).

Définition 3.6 (Taux de symétrie) *Le taux de symétrie S est défini comme le pourcentage d'interactions détectées de manière symétrique par rapport à l'ensemble des interactions entre des protéines VBP :*

$$S = 100 \times \frac{|Interactions_{\text{Réciproques}}|}{|Interactions_{Total}|} \tag{3.2}$$

Nous avons remarqué que les jeux de données restreints avaient un taux de symétrie meilleur que les jeux de données d'origine (voir la Table 3.2). Ceci vient confirmer l'idée que ces jeux de données restreints sont de meilleure qualité. De plus, nous avons noté des différences importantes entre les jeux de données, puisque le taux de symétrie variaient de 5,2% à 40,7%. La meilleure valeur a été obtenue pour l'ensemble des interactions de catégories A extraites de *RainFull*.

3.1.4.2 Asymétrie au niveau des protéines

Après avoir quantifié l'asymétrie au niveau des interactions, nous avons voulu quantifier l'asymétrie au niveau des protéines. En effet, si la technologie était parfaite, chaque interaction serait détectée dans les deux sens. Ainsi, pour une protéine donnée, les degrés entrants et sortants seraient les mêmes. Nous savons que ce n'est pas le cas, d'après les résultats précédents. C'est pourquoi nous avons choisi d'étudier les degrés d'une protéine donnée, de manière à quantifier l'asymétrie observée pour cette protéine. Nous nous sommes concentrés sur les interactions non symétriques pour voir si ces interactions étaient biaisées dans un sens ou dans l'autre, c'est-à-dire si cette asymétrie a tendance à être plutôt dans un sens ou pas. Là encore, nous avons considéré seulement les interactions entre des protéines VBP, pour les mêmes raisons que précédemment (voir Section 3.1.4.1).

Une protéine peut détecter un certain nombre de proies, qui est représenté par son degré sortant. De plus, cette même protéine peut avoir été détectée par un certain nombre d'appâts, qui est représenté par son degré entrant. Ici, nous considérons les degrés non symétriques.

- n_{out} : nombre de proies détectées par la protéine, ne l'ayant pas détectée
- n_{in} : nombre d'appâts ayant détecté la protéine, non détectés par elle

Dans un premier temps, nous avons représenté chaque protéine en fonction de ses degrés sortant n_{out} et entrant n_{in}. En d'autres termes, chaque protéine est représentée en fonction du nombre de proies qu'elle a détectées et du nombre d'appâts qui l'ont détectée elle (voir Figure 3.5). Les protéines n'ayant que des interactions symétriques sont à l'origine ; les protéines ayant autant d'interactions non symétriques dans un sens que dans l'autre, sont sur la diagonale ; les autres protéines ont un biais de symétrie.

Nous avons alors appliqué le modèle binomial proposé par Chiang *et al.* afin de quantifier l'asymétrie liée à une protéine donnée [Chiang *et al.*, 2007]. Nous avons ainsi identifié 13 protéines VBP montrant un biais systématique en fixant un seuil de confiance à 10^{-3} (voir Annexe H). Cette proportion de protéines montrant un biais systématique est inférieure à 3% comme observé par Chiang *et al.* [Chiang *et al.*, 2007]. Il est intéressant de noter qu'aucune protéine n'a été détectée avec un tel biais systématique pour le même seuil dans le graphe *SatoCore*. Ceci confirme à nouveau la qualité des jeux de données restreint.

Les protéines qui ont un degré entrant différent de leur degré sortant peuvent être séparées en deux groupes : d'un côté les protéines dont le degré entrant est supérieur au degré sortant. Ces protéines ont tendance à être détectées par un grand nombre d'appâts (*e. g.* proie collante), ou ne permettent pas de détecter les proies (appât qui fonctionne mal). D'un autre côté, les autres protéines ont un degré entrant inférieur au degré sortant. Dans ce cas, elles ont tendance à identifier beaucoup de proies (*e. g.* appât collant), ou à ne pas être détectées comme proie (proie inefficace).

Pour quantifier ces phénomènes, nous avons calculé un score associé à chaque protéine, indiquant l'excès des degrés entrants par rapport aux degrés sortants [Chiang *et al.*, 2007]. Ces scores ont été calculés selon l'équation 3.3. Ainsi, un score positif caractérise une protéine du premier groupe (proie collante), alors qu'un score négatif caractérise une protéine du second groupe (proie inefficace). Nous avons alors représenté la densité des scores pour les différents jeux de données (voir Figure 3.6 et Annexe H).

$$z = \frac{n_{in} - n_{out}}{\sqrt{n_{in} + n_{out}}} \tag{3.3}$$

Pour les données issues des approches double-hybrides (*Y2H*), les profils sont globalement similaires à l'exception de *ItoFull*. Ce profil montre une moyenne positive. Ceci peut s'expliquer par la présence d'un certain nombre d'appâts auto-activateurs qui augmentent artificiellement le degré entrant d'un certain nombre de protéines, augmentant par suite leur score.

Pour les données issues des approches *AP-MS*, les profils montrent une moyenne négative. Ceci peut s'expliquer en considérant l'abondance des protéines dans les conditions d'expérience. Prenons par exemple une protéine p faiblement exprimée, qui est taguée et exprimée comme appât. Supposons qu'elle détecte k proies $p_1, p_2, ..., p_k$. La démarche inverse pour chacune des proies a une probabilité plus faible de trouver p, de par son faible niveau d'expression. De plus, même si la protéine p peu abondante est descendue par le pull-down, il se peut que la spectrométrie de masse ne permette pas de détecter p à l'intérieur du mélange de protéines. Pour ces deux raisons, l'approche *AP-MS* est

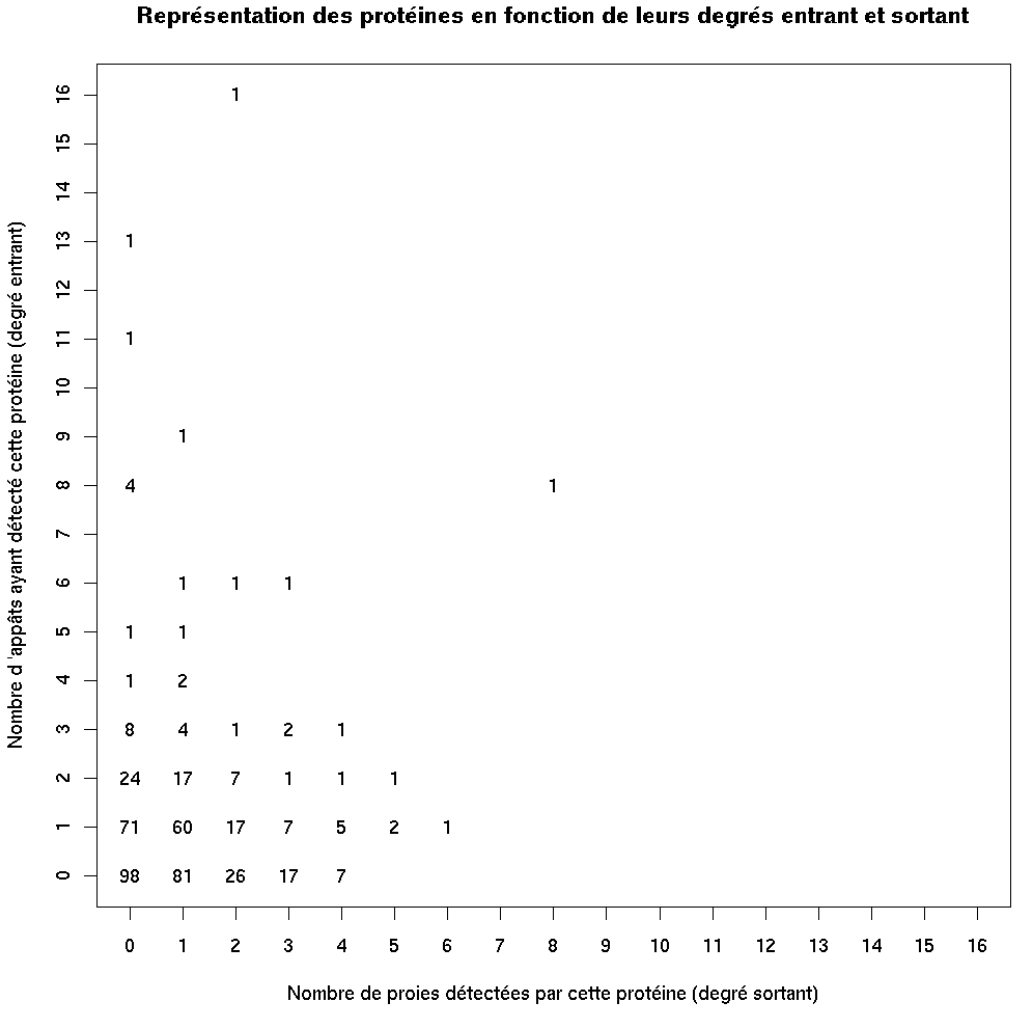


FIG. 3.5 – **Degrés entrant et sortant non réciproques des protéines VBP entre elles pour l'étude de Sato *et al.*** Chaque protéine est représentée en fonction du nombre de proies qu'elle a détectées, et du nombre d'appâts qui l'ont détectée elle. Les nombres indiquent combien de protéines ont tel degré entrant (ordonnées) et tel degré sortant (abscisses). Seules les interactions non symétriques ayant lieu entre des protéines VBP sont prises en compte. Les protéines n'ayant que des interactions symétriques sont à l'origine ; les protéines ayant autant d'interactions non symétriques dans un sens que dans l'autre, sont sur la diagonale ; les autres protéines ont un biais de symétrie.

plus sensible dans la détection des membres d'un complexe pour un appât particulier que dans le sens inverse [Chiang *et al.*, 2007].

Cette représentation permet de refléter des biais qui mériteraient d'être explicités de manière plus précise et détaillée.

Les interactions identifiées par Sato *et al.* sont donc assez proches des précédents jeux de données obtenus par *Y2H* en termes de couverture de protéines et d'interactions mais également en termes de biais systématique provenant de l'asymétrie de la technique d'identification. Le jeu de données restreint *SatoCore* est largement moins dense que le jeux de données complet *SatoFull* et de meilleure qualité comme l'ont montré le meilleur taux de symétrie et l'absence de protéines possédant un biais systématique.
Après avoir analysé ce jeu de données expérimental, l'idée a été de le comparer avec les interactions prédites chez *Synechocystis* par les deux méthodes de prédiction d'interactions protéine-protéine que nous avons développées (voir Chapitre 2).

3.2 Comparaison des listes d'interactions entre les données expérimentales et prédites

Nous avons tout d'abord voulu identifier les interactions communes aux jeux de données prédits et expérimentaux. Pour cela, nous devions avoir une modélisation commune pour les deux approches, c'est pourquoi nous avons commencé par adapter les données expérimentales, afin d'identifier l'intersection des deux jeux de données en termes d'interactions.

3.2.1 Identification de l'intersection

Les méthodes de prédiction que nous avons développées nous ont permis de construire un réseau d'interactions protéine-protéine pour *Synechocystis* constitué de 8 783 interactions entre 1 011 protéines et dénommé *InteroFull* (voir Chapitre 2). Les interactions expérimentales et prédites sont marquées par deux différences techniques principales.

À la différence des interactions identifiées expérimentalement, les interactions prédites *in-silico* sont symétriques. Ainsi, la notion d'appât et de proie n'existe pas. Il est vrai que ces interactions proviennent du transfert d'interactions identifiées expérimentalement, dont la plupart ont effectivement été identifiées par des techniques asymétriques. Par conséquent, ces interactions prédites peuvent refléter les biais présents dans les jeux de données sources. Néanmoins, elles ne sont pas orientées à cause de la gestion des bases de données. En effet, la plupart des bases de données ne gardent pas les interactions sous forme orientées. De plus, certaines techniques expérimentales ne donnent pas lieu à des observations orientées.

Par ailleurs, le processus d'inférence étant basé sur les numéros d'accession de la base de données Uniprot, toute protéine non identifiée dans Uniprot n'a pas pu être prise en

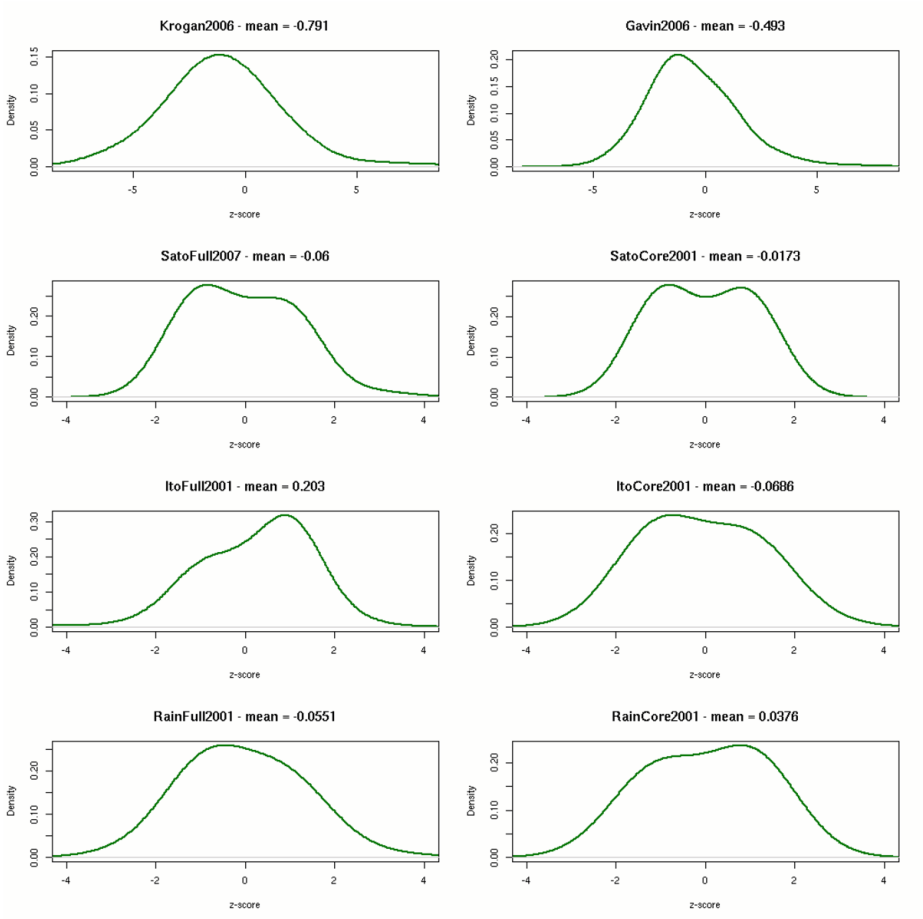


FIG. 3.6 – **Distribution des scores sur l'ensemble des protéines VBP pour quelques études à grande échelle.** Chaque graphique représente la densité du score pour le jeu de données dont le nom est indiqué au-dessus, associé à la valeur moyenne du score. Seules les interactions entre des protéines VBP sont considérées. Ce score permet d'évaluer la tendance des protéines à être une proie collante ou mal détecter les proies (score positif), ou encore à être un appât collant ou à être mal détectée (score négatif). Plus de détails sur le score sont donnés dans l'Annexe H.

compte lors de la construction *in-silico* du réseau d'interactions protéine-protéine chez *Synechocystis*.

Afin de mener une comparaison pertinente entre les interactions expérimentales et prédites, nous avons créé un nouveau jeu de données issu de *SatoFull*. Ce jeu de données, dénommé *SatoFull_Uni_Bi*, ne contient que des interactions non orientées distinctes. Pour cela, nous avons regroupé les interactions orientées réciproques (A→B et B→A dans *SatoFull* deviennent A-B dans *SatoFull_Uni_Bi*). De plus, les interactions impliquant au moins une protéine absente de la base de données Uniprot ont été retirée. Le nouveau jeu de données *SatoFull_Uni_Bi* est ainsi constitué de 2 970 interactions entre 1 846 protéines.

Les jeux de données *InteroFull* et *SatoFull_Uni_Bi* possèdent 25 interactions en commun ayant lieu entre 40 protéines (voir Figure 3.7). Comme nous l'avons précédemment rappelé (voir Section 2.3.5), il a été montré que les jeux de données issus d'identifications à haut-débit d'interactions protéine-protéine ne se recoupent que faiblement. En effet, moins de 10% du nombre total d'interactions sont retrouvées lorsque la même technique est utilisée chez la même espèce [Arifuzzaman *et al.*, 2006]. Ceci souligne le fort taux de faux négatifs de ces techniques, ce que nous retrouvons également ici lors de la comparaison avec les prédictions. Pour évaluer la pertinence de ce recouvrement entre les interactions prédites et identifiées expérimentalement, nous avons calculé la probabilité de trouver par hasard un recouvrement au moins aussi grand que celui observé. Nous avons utilisé pour cela un modèle hypergéométrique, selon lequel la probabilité est inférieure à $2,0 \times 10^{-5}$ (voir la section C.5 de l'annexe C). Par conséquent, les données expérimentales corroborent les prédictions.

3.2.2 Description de l'intersection

Les 25 interactions prédites *in-silico* et identifiées expérimentalement impliquent 40 protéines. Parmi ces 25 interactions, nous avons identifié des complexes stables comme le ribosome ou l'ARN polymérase (voir Figure 3.7). De plus, nous avons observé également des interactions entre des protéines non connues pour former des complexes, comme par exemple les enzymes carboxylases. Il peut s'agir alors d'interactions plus transitoires.

Par définition, l'ensemble de ces interactions est caractérisé à la fois par le processus d'inférence et par l'identification expérimentale.

Concernant l'inférence, nous avons noté que 24 de ces interactions étaient prédites par la méthode *InteroBH*, six par la méthode *InteroPorc*, dont cinq étaient prédites par les deux méthodes. Rappelons que les interactions prédites *in-silico* ont été transférées à partir de sept espèces sources (voir Section 2.1). Même si la plupart des interactions prédites et identifiées expérimentalement ont été transférées depuis *Escherichia coli* (72%), il est intéressant de remarquer que six des sept espèces sources ont été utilisées. En effet, seul *Caenorhabditis elegans* n'a pas permis de transférer des interactions qui font partie de *SatoFull_Uni_Bi*. De plus, ces prédictions ne sont pas redondantes pour la plupart. Ainsi, si on retire une des espèces sources parmi les cinq suivantes (*E. coli*, *S. cerevisiae*, *D. melanogaster*, *A. thaliana* et *H. sapiens*), on perd au moins une interaction. Enfin, nous avons calculé le nombre d'interactions identifiées expérimentalement pour les trois

Graphes	Interactions	Protéines	VB	VP	VBP
SatoCore	1 064	1 152	631	716	195
SatoFull	3 236	1 920	1 044	1 352	476
Core/Full	**33%**	**60%**	**40%**	**53%**	**41%**
ItoCore	839	795	455	504	164
ItoFull	4 524	3 242	1 522	2 493	773
Core/Full	**19%**	**25%**	**29%**	**20%**	**21%**
RainCore	622	509	232	397	120
RainFull	1 568	740	256	632	148
Core/Full	**40%**	**69%**	**91%**	**63%**	**81%**

TAB. 3.1 – **Comparaison entre les jeux de données *FULL* et *CORE* pour trois approches par *Y2H*.** La colonne **Graphes** contient les noms des jeux de données. Les colonnes **Interactions** et **Protéines** contiennent respectivement le nombre d'interactions et de protéines de ce jeux de données. Enfin les colonnes **VB**, **VP** et **VBP** indiquent respectivement le nombre d'appâts viables, de proies viables et d'appâts/proies viables. Pour chaque étude, nous avons indiqué le pourcentage relatif des jeux de données restreints par rapport aux jeux de données complets. Notons que les rapports entre *ItoCore* et *ItoFull* sont assez différents des deux autres jeux de données, qui sont proches en termes de rapports entre les nombres de protéines et d'interactions, mais assez différents pour ce qui concerne la répartition des protéines.

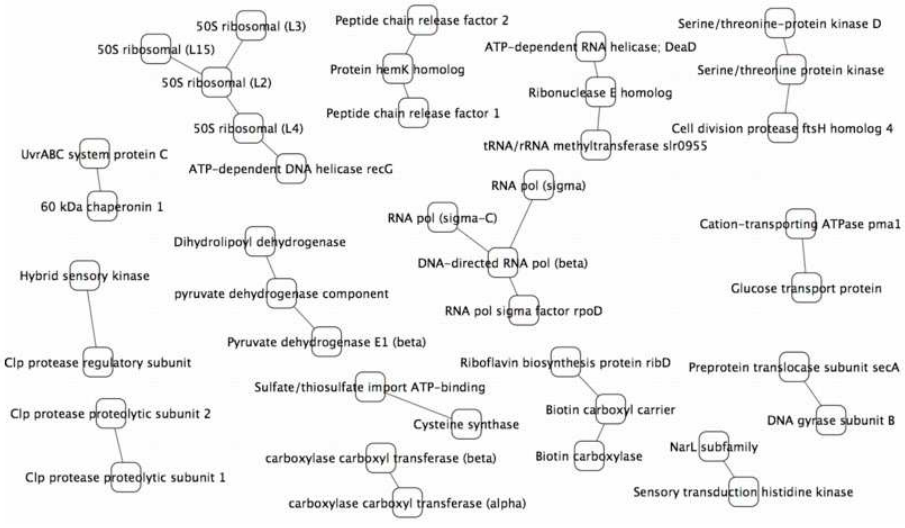


FIG. 3.7 – **Interactions prédites et identifiées expérimentalement.** Chacun des 40 nœuds représente une protéine de *Synechocystis* associée à sa description fournie par Uniprot pour décrire sa (principale) fonction. Les arêtes entre les nœuds représentent des interactions prédites par nos méthodes de prédiction et identifiées expérimentalement par Sato *et al.* [Sato *et al.*, 2007].

Graphes	**Protéines**	**Interactions**	**Réciproques**	**Taux de symétrie**
SatoCore	131	61	12	19,7%
SatoFull	406	439	23	5,2%
Core/Full				**× 3,8**
ItoCore	128	115	36	31,3%
ItoFull	732	863	75	8,7%
Core/Full				**× 3,6**
RainCore	86	68	24	35,3%
RainFull	120	169	26	15,4%
Core/Full				**× 2,3**
SatoA	80	27	5	18,5%
SatoFull	406	439	23	5,2%
A/Full				**× 3,5**
RainA	79	59	24	40,7%
RainFull	120	169	26	15,4%
A/Full				**× 2,6**

TAB. 3.2 – **Comparaison des taux de symétrie entre les jeux de données *FULL* et *CORE*.** Pour ces analyses, les interactions ont été filtrées. Seules les interactions entre deux protéines VBP différentes ont été considérées. La colonne **Graphes** contient les noms des jeux de données. Les colonnes **Protéines** et **Interactions** contiennent respectivement le nombre de protéines et d'interactions considérées. La colonne **Réciproques** indique le nombre d'interactions qui ont été détectées dans les deux sens (de manière réciproque). Enfin la dernière colonne indique la valeur du taux de symétrie, défini comme le pourcentage d'interactions détectées de manière réciproque parmi toutes les interactions mettant en jeux deux protéines VBP. Pour chaque étude, nous avons indiqué le rapport entre ces deux valeurs de ratio de symétrie pour les jeux de données restreints, par rapport aux jeux de données complets.

réseaux prédits par la méthode *InteroBH*, à trois niveaux de confiance différents (voir Figure 3.8). Nous avons observé que le pourcentage d'interactions identifiées expérimentalement parmi les prédictions augmentait lorsque le seuil de confiance augmentait. Par conséquent, ceci soutient la pertinence du paramètre choisi pour évaluer la qualité des prédictions.

En ce qui concerne l'identification expérimentale, 11 interactions sont assignées aux catégories A et B (donc dans *SatoCore*) ; 12 interactions sont assignées à la catégorie C, et trois interactions à la catégorie D. De plus, nous nous sommes intéressés aux protéines conservées chez *A. thaliana* dans la mesure où Sato *et al.* ont sélectionné, pour une partie des appâts, des protéines de *Synechocystis* ayant des homologues chez *A. thaliana*. Une seule interaction a été transférée à partir d'*A. thaliana*. Ceci peut s'expliquer par le fait que très peu de prédictions ont été faites à partir d'*A. thaliana* car les données sources étaient très limitées (1 466 interactions chez *A. thaliana* dans la base de données IntAct en Avril 2007).

Nous avons identifié et analysé 25 interactions prédites *in-silico* par nos méthodes de prédiction, et également mises en évidence expérimentalement par Sato *et al.*
Au-delà des simples listes d'interactions, l'idée a ensuite été de s'intéresser à l'organisation des protéines les unes avec les autres.

3.3 Comparaison des topologies entre les réseaux expérimentaux et prédits

Nous avons voulu comparer les structures des réseaux prédits et expérimentaux. Pour cela, nous avons étudié différents paramètres topologiques des graphes associés, afin de caractériser certaines propriétés locales et/ou globales. Les définitions des différents paramètres sont données dans l'annexe H.

3.3.1 Analyse des paramètres globaux

Tout d'abord, nous avons étudié la structure globale des graphes en calculant, outre les nombres de nœuds et d'arêtes, le diamètre, le degré moyen, le coefficient de clustering moyen et la taille de la plus grande composante connexe (voir Table 3.3). Ces paramètres permettent d'évaluer entre autres la densité et le degré de connectivité des graphes. Ainsi, nous avons noté une importante différence, en termes de densité, entre les réseaux prédits et expérimentaux. En effet, les réseaux expérimentaux sont beaucoup moins denses que les réseaux prédits, ce qui se traduit par un diamètre plus élevé et un degré moyen plus faible. De plus, le réseau *SatoCore* est beaucoup moins connecté que les autres, qui sont principalement réduits à une grande composante connexe.

D'autre part, le coefficient d'agglomération, ou coefficient de clustering, permet d'évaluer la vraisemblance que deux nœuds voisins d'un nœud donné soient connectés entre eux. Pour un graphe, le coefficient de clustering moyen est calculé en considérant la

moyenne arithmétique des coefficients de clustering de tous les nœuds du graphe. Un coefficient élevé indique une forte modularité. Il a été montré [Watts et Strogatz, 1998] que les graphes aléatoires ont un coefficient de clustering moyen proche du rapport entre le degré moyen et le nombre de nœuds (voir Équation H.2). Tous les graphes considérés ici, aussi bien prédits qu'expérimentaux, ont un coefficient de clustering moyen largement supérieur à cette approximation, montrant leur forte modularité. Ceci était attendu car c'est le cas pour la plupart des réseaux d'interactions protéine-protéine. Néanmoins, les graphes expérimentaux et prédits montrent des degrés de modularité différents.

$$\overline{CC} \approx \frac{\overline{k}}{N} \tag{3.4}$$

Ainsi, les graphes considérés, provenant aussi bien de la prédiction *in-silico*, que de la détection expérimentale, sont fortement modulaires. En revanche, les réseaux expérimentaux sont beaucoup moins denses que les réseaux prédits. Pour aller vers une analyse plus détaillée, nous nous sommes intéressés à l'évolution de certains de ces paramètres en fonction du degré des protéines considérées, c'est-à-dire à la distribution de ces paramètres.

3.3.2 Analyse des distributions des coefficients

Après les propriétés globales, nous nous sommes intéressés à des propriétés plus locales en étudiant les distributions de deux paramètres : le coefficient de clustering, qui mesure la proportion de voisins connectés, et le coefficient de voisinage, encore appelé coefficient topologique, évalue dans quelle mesure un nœud partage des voisins avec les autres nœuds(les définitions sont précisées dans l'annexe H.2). Ces distributions indiquent la moyenne du paramètre étudié pour tous les nœuds de degré k, k prenant toutes les valeurs possible selon le réseau étudié. Elles ont été calculées en utilisant le plugin NetworkAnalyzer [Assenov *et al.*, 2007] du logiciel Cytoscape [Shannon *et al.*, 2003].

Concernant le coefficient de clustering, les distributions sont largement différentes pour *SatoFull* et *InteroFull* (voir Figure 3.9). Le réseau expérimental *SatoFull* étant moins modulaire, les valeurs ne sont pas du même ordre de grandeur. De plus, dans le cas du graphe *InteroFull*, la distribution est décroissante. Ceci indique que le graphe a une nature hiérarchique et suggère qu'il possède deux niveaux d'organisation : le clustering local représente les modules fonctionnels, et la connectivité globale peut être vue comme des points de communication entre ces complexes [Stelzl *et al.*, 2005]. En revanche, ce n'est pas du tout le cas pour le graphe *SatoFull*.

Le coefficient de voisinage évalue quant à lui dans quelle mesure un nœud partage des voisins avec les autres nœuds. Pour *SatoFull* et *InteroFull*, la distribution est décroissante (voir Figure 3.10). Ceci montre que les protéines à fort degré n'ont pas plus de partenaires en commun avec l'ensemble des protéines, que ne l'ont les protéines faiblement connectées.

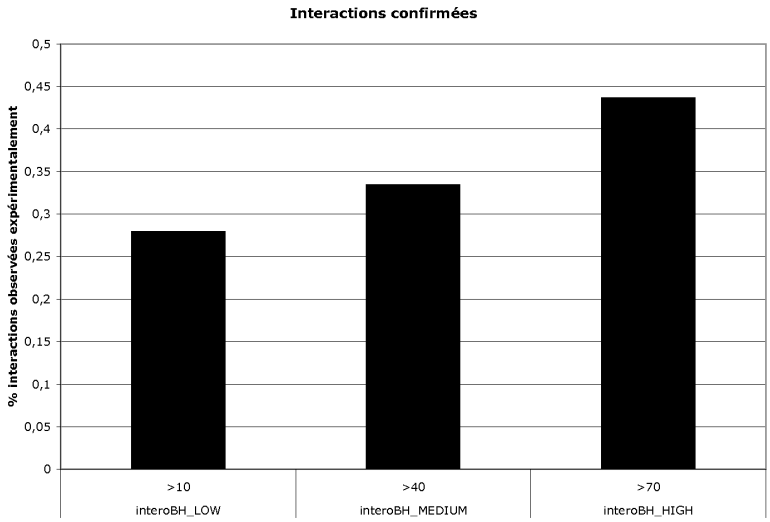


FIG. 3.8 – **Pourcentage d'interactions identifiées expérimentalement parmi les prédictions obtenues par *InteroBH*.** Cet histogramme indique le pourcentage d'interactions identifiées expérimentalement parmi les interactions prédites par interoBH pour trois niveaux de confiance différents. Le niveau de confiance est défini par la E-value jointe (voir Section 2.1.1). La valeur indiquée pour chacun des trois réseaux *InteroBH_LOW*, *InteroBH_MEDIUM* et *InteroBH_HIGH*, est le score associé à la E-value jointe ($S = -log(Evalue)$).

Réseaux	**M**	**N**	**D**	$\overline{\mathbf{k}}$	**Paires connectées**	
					Nb	**%**
SatoFull	3 236	1 920	16	3,2	3 372 833	91
SatoCore	1 064	1 152	26	1,7	319 136	24
InteroFull	8 783	1 011	6	17,4	1 017 074	99
InteroPorc	1 446	384	8	7,5	139 520	94
InteroBH_HIGH	2 748	741	6	7,4	548 340	100
InteroBH_MEDIUM	5 070	884	7	11,5	777 044	99
InteroBH_LOW	8 586	998	6	17,2	991 022	99

TAB. 3.3 – **Propriétés topologiques des réseaux prédits et expérimentaux.** Pour chaque réseau sont indiqués les paramètres suivants : le nombre d'arêtes (interactions) (**M**), le nombre de nœuds (protéines) (**N**), le diamètre (**D**), le degré moyen ($\overline{\mathbf{k}}$), le nombre de paires connectées (**Nb**), et le pourcentage qu'elles représentent par rapport à l'ensemble des paires de protéines (**%**).

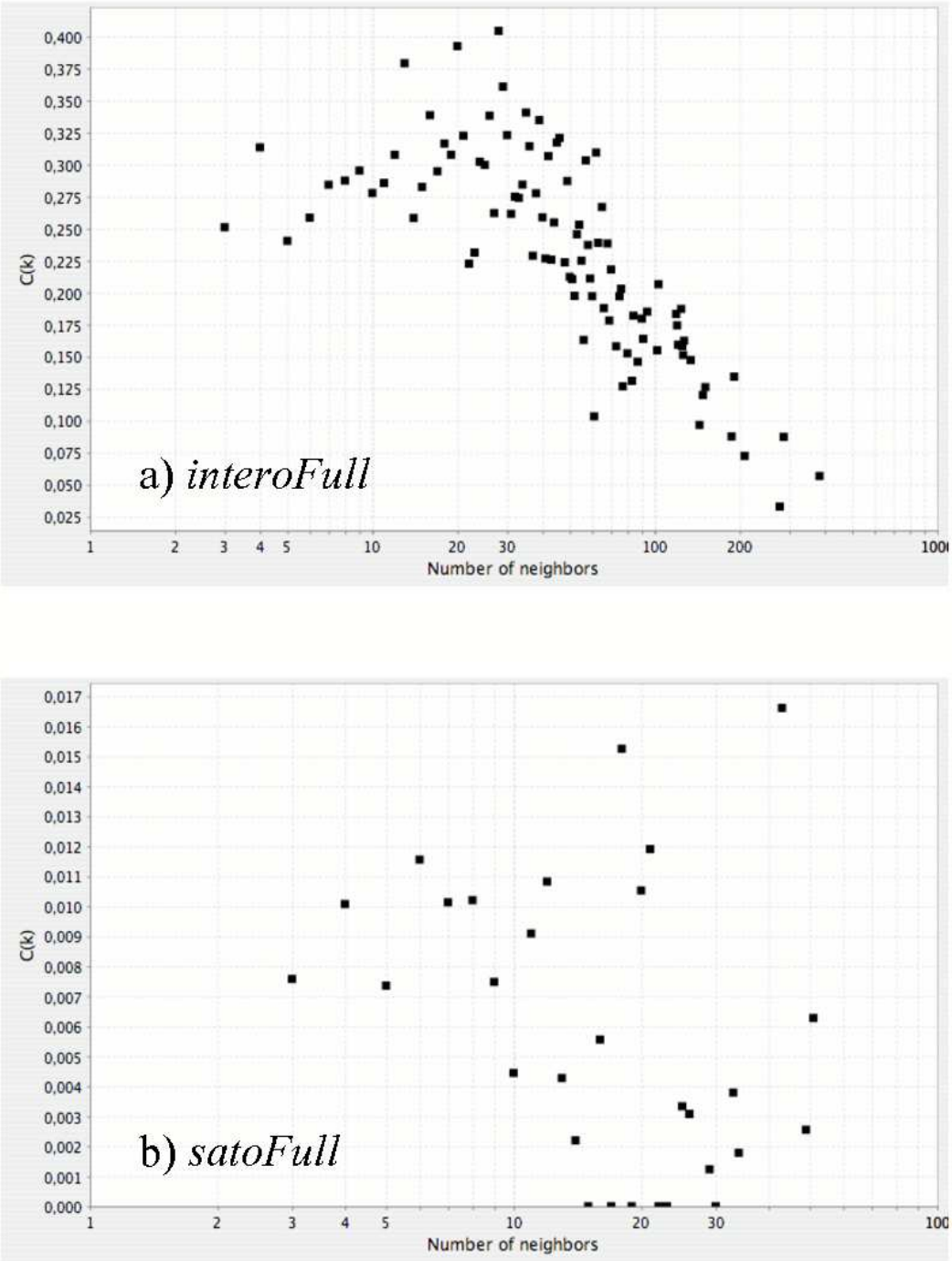


FIG. 3.9 – **Coefficient de clustering pour les réseaux prédits et expérimentaux.** Les graphiques montrent les distributions du coefficient de clustering pour : a) *InteroFull*, le réseau prédit ; b) *SatoFull*, le réseau expérimental de Sato *et al.*Il est important de noter la différence entre les deux échelles pour les valeurs du coefficient de clustering.

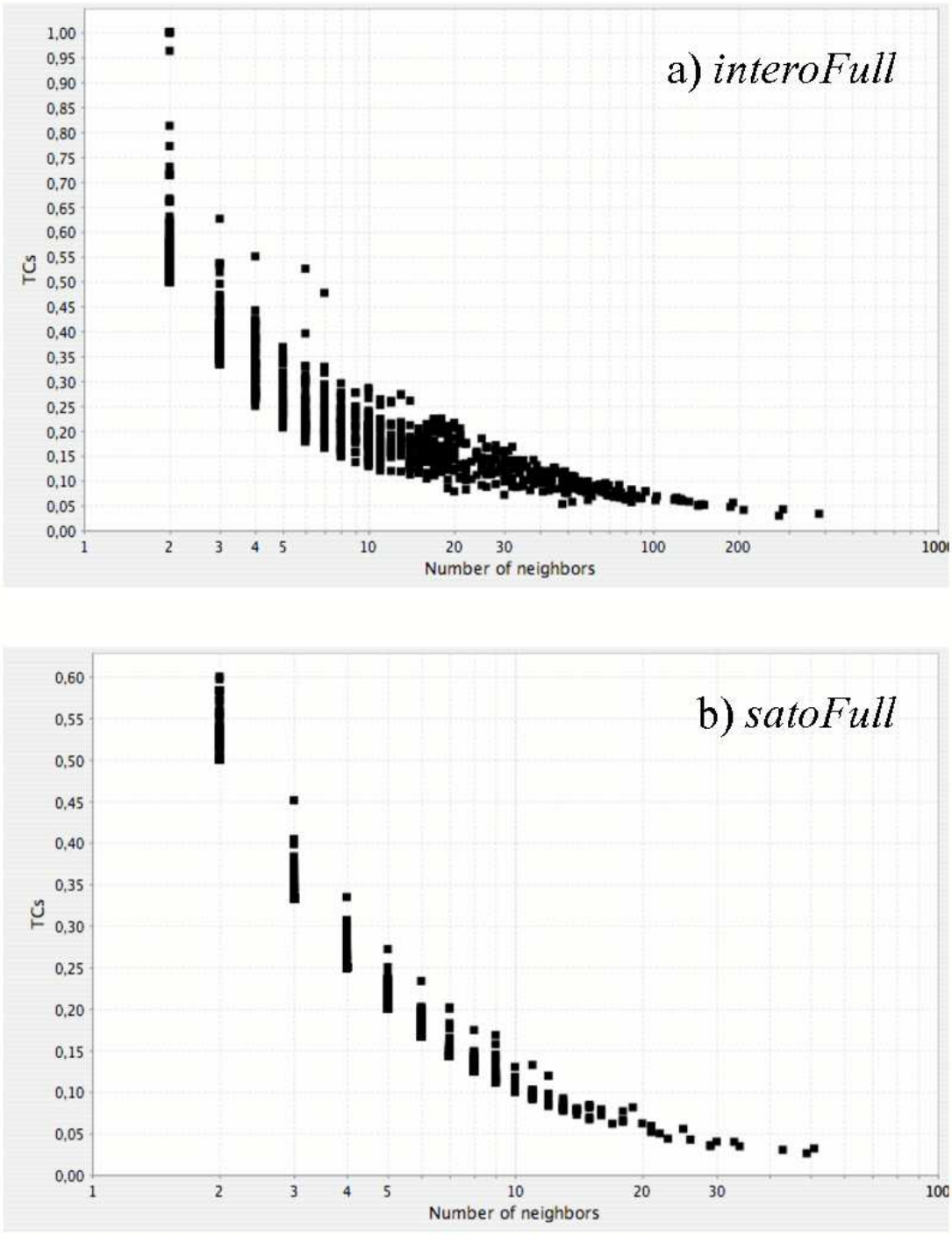


FIG. 3.10 – **Coefficient de voisinage pour les réseaux prédits et expérimentaux.** Les graphiques montrent les distributions du coefficient de voisinage pour : a) *InteroFull*, le réseau prédit ; b) *SatoFull*, le réseau expérimental de Sato *et al.*

Ainsi, même si les réseaux expérimentaux et prédits sont modulaires, *InteroFull* est malgré tout beaucoup plus clusterisé que *SatoFull*. Ceci est cohérent avec la précédente conclusion, selon laquelle *InteroFull* est beaucoup plus dense que *SatoFull*. Les réseaux expérimentaux et prédits montrent donc certains aspects similaires et d'autres différents. L'idée a alors été de rechercher un modèle de graphe pour lequel chacun des réseaux s'adapterait bien, dans le but de quantifier les différentes caractéristiques des réseaux.

3.3.3 Recherche d'un modèle de graphe aléatoire

Pour rechercher un modèle de graphe aléatoire qui s'adaptait bien à un réseau donné, nous avons comparé ce réseau réel donné avec un certain nombre de graphes aléatoires suivant plusieurs modèles, en considérant à la fois les propriétés globales et locales. Pour cela, nous avons utilisé les cinq modèles de graphes aléatoires suivants :

1. modèle Erdös-Rényi (*er*) : dans le modèle $G(N, M)$, un graphe est choisi aléatoirement parmi tous les graphes de N nœuds et M arêtes ;
2. modèle Erdös-Rényi avec la même distribution des degrés (*er_ dd*) : il s'agit d'un graphe aléatoire d'Erdös-Rényi avec une distribution des degrés fixée ;
3. modèle géométrique de dimension n (*geo*) : il s'agit d'un ensemble de points dans un espace métrique qui sont reliés entre eux s'ils sont suffisamment proches ;
4. modèle invariant d'échelle ou scale-free (*sf*) : un graphe est dit invariant d'échelle si sa distribution des degrés peut s'exprimer sous la forme $p(k) = k^{-\gamma}$;
5. modèle sticky (*sticky*) : ce modèle a été développé pour rendre compte des interactions protéine-protéine ; une arête est créée entre deux nœuds selon leur tendance à être liés à d'autres nœuds (indice de *stickiness*).

Les simulations ont été réalisées avec le logiciel GraphCrunch [Milenkovic *et al.*, 2008]. L'intérêt de ce logiciel est notamment de considérer, en plus des propriétés globales relativement classiques, des propriétés locales concernant la nature des graphlets contenus dans les graphes comparés. Les graphlets sont des petits sous-graphes induits. Ainsi, nous avons évalué la différence entre les fréquences relatives des graphlets en calculant la RGF-distance (*Relative Graphlet Frequency Distance*). De plus, nous avons évalué la ressemblance entre les distributions des degrés des graphlets en calculant le GDD-agreement (*Graphlet Degree Distribution Agreement*). Le principe général est le suivant :

1. les propriétés du réseau réel donné en entrée sont calculées (nombre de nœuds N, nombre d'arêtes M, distribution des degrés) ; les réseaux réels sont ici les graphes expérimentaux et prédits ;
2. un certain nombre d'instances de graphes sont créées en suivant chacun des modèles aléatoires et en conservant les nombres de nœuds N et d'arêtes M du réseau réel, ainsi que la distribution des degrés pour le modèle *er_ dd* (voir l'annexe H.3 pour la génération des graphes) ; nous avons générés 50 graphes aléatoires pour chacun des modèles ;
3. les propriétés globales (diamètre moyen, distribution des coefficients de clustering, distribution des degrés, distribution des plus courts chemins) et locales (fréquences

relatives des graphlets, distribution des degrés des graphlets) de chacun des graphes générés sont comparées avec celles du réseau réel grâce à des mesures de similarité ;

4. pour chaque modèle de graphes, la moyenne et l'écart-type des similarités sont calculés sur l'ensemble des instances de graphes générés suivant ce modèle.

Les résultats sont illustrés pour le graphe réel *InteroFull*. Les paramètres globaux montrent une bonne concordance avec les modèles er_dd et sticky (voir Figure H.5) ; les paramètres locaux, et en particulier les fréquences des graphlets, montrent une meilleure concordance avec le modèle *sicky* (voir Figure H.6). Ainsi, ces simulations ont permis de conclure que le modèle *sticky* est celui qui semble s'adapter le mieux aux graphes prédits. Ceci n'est pas très étonnant dans la mesure où ce modèle a été développé pour rendre compte spécifiquement des interactions protéine-protéine, en se basant sur la présence de surfaces d'interaction ou de domaines de liaison [Przulj et Higham, 2006]. De plus, il faut noter que le programme GraphCrunch a été développé par l'équipe qui a défini ce modèle de graphe.

Concernant les graphes expérimentaux, *SatoFull* est aussi plus proche du modèle *sticky* (voir l'annexe H.3). En revanche, plusieurs modèles semblent s'adapter au graphe *SatoCore*, en particulier le modèle scale-free. Il est intéressant de noter que ces deux jeux de données (*FULL* et *CORE*) ne semblent pas s'adapter au même modèle. Ce phénomène a déjà été observé chez *D. melanogaster* [Milenkovic *et al.*, 2008], soulignant les différences entre un réseau de haute confiance mais moins complet, et un réseau plus grand mais plus bruité. Il serait intéressant de tester d'autres modèles de graphes plus évolués, comme par exemple des mélanges d'Erdös-Rényi [Picard *et al.*, 2006].

Nous avons donc montré que les réseaux expérimentaux et prédits présentent certains points de ressemblance. En particulier, ils ont un caractère modulaire, caractéristique des réseaux dits réels comme les réseaux sociaux et les réseaux internet, et à la différences des réseaux aléatoires. De plus, le modèle de graphe *sticky* est celui qui s'adapte le mieux aux graphes aussi bien expérimentaux que prédits. En revanche, les réseaux prédits sont beaucoup plus denses que les réseaux expérimentaux et plus clusterisés.
À cette étape, l'idée a alors été d'extraire de ces réseaux les modules fonctionnels caractérisés par des sous-graphes denses.

3.4 Comparaison des décompositions en modules

Depuis quelques années, des méthodes de clustering ont été développées et appliquées à différents réseaux d'interactions protéine-protéine afin d'en extraire des modules fonctionnels, comme des complexes protéiques [Poyatos et Hurst, 2004] ou des réseaux de régulation [Rives et Galitski, 2003]. Depuis l'article de Hartwell *et al.* sur la modularité des réseaux [Hartwell *et al.*, 1999], un très grand nombre de définitions ont été proposées pour décrire les modules. Ici, nous avons voulu extraire les modules fonctionnels pour chacun des réseaux étudiés en considérant les sous-graphes denses. Pour cela, nous avons

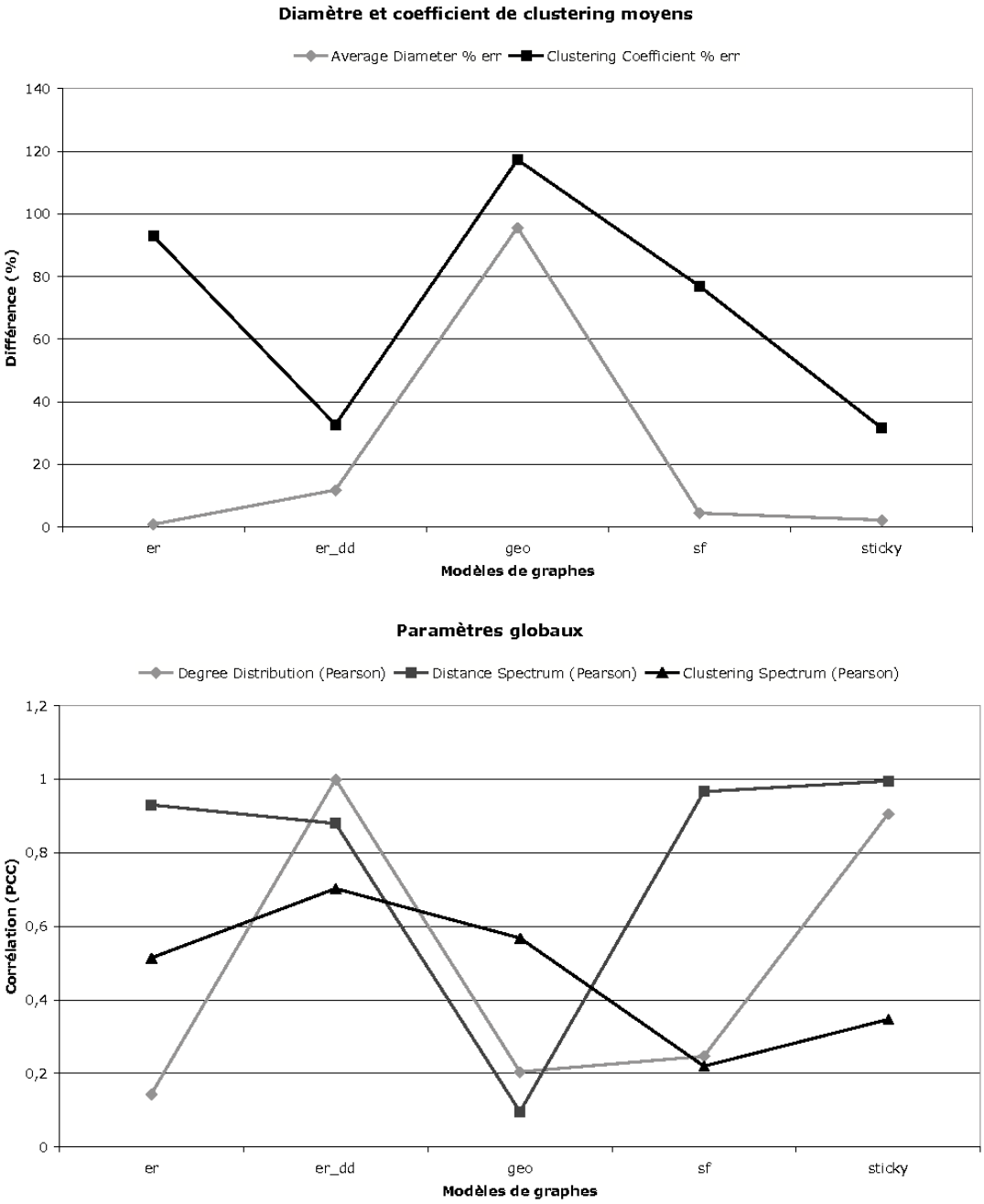


FIG. 3.11 – **Comparaison des paramètres globaux.** Cette figure illustre la comparaison des paramètres globaux du réseau réel *InteroFull* et de 50 instances de chacun des modèles aléatoires (*er*, *er_ dd*, *geo*, *sf*, *sticky*). Le graphe du haut montre la différence en pourcentage entre les diamètres moyens, et également entre les coefficients de clustering moyens. Le graphe du bas montre les coefficients de corrélation entre les distributions des degrés, les distributions des plus courts chemins et les distributions des coefficients de clustering.

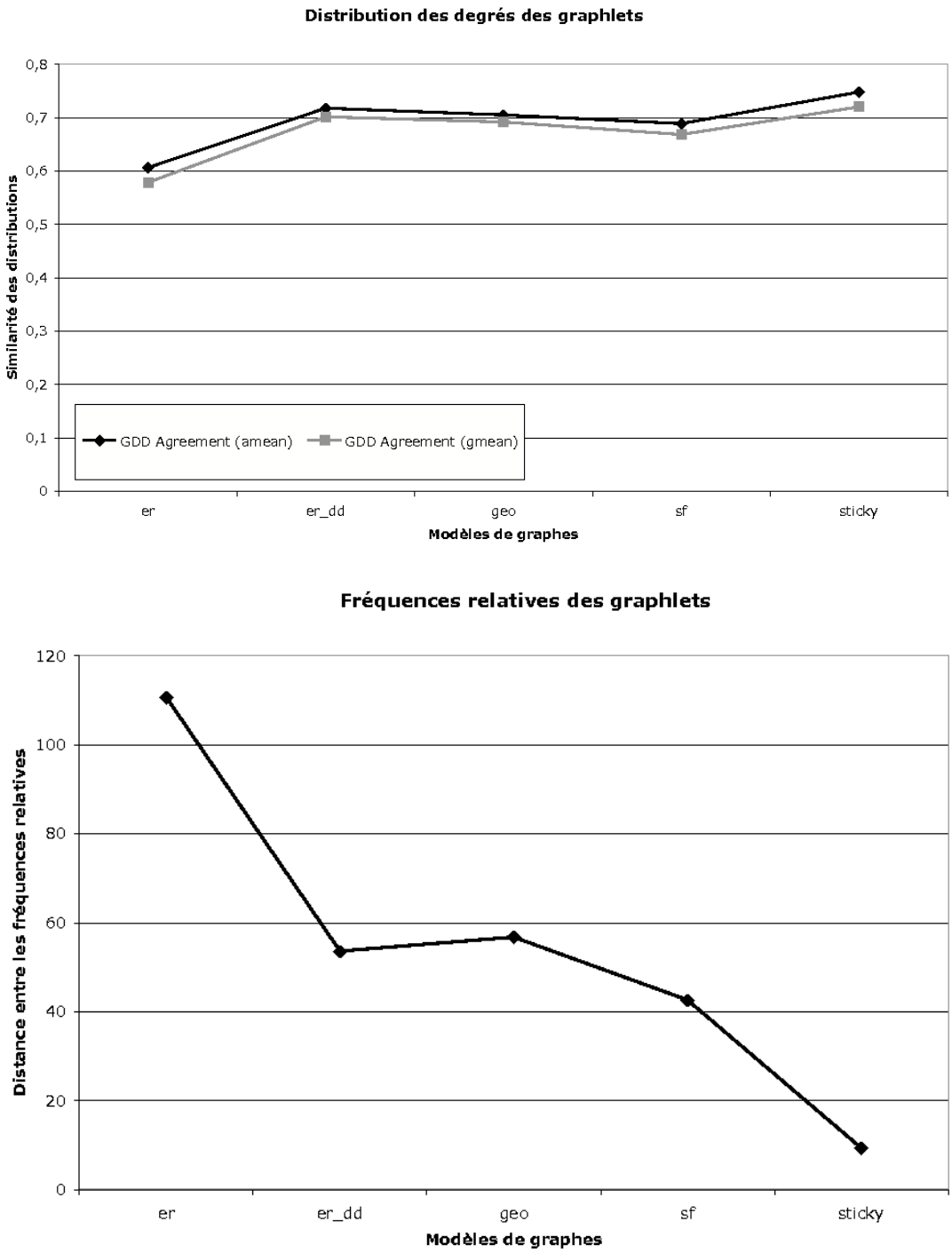


FIG. 3.12 – **Comparaison des paramètres locaux.** Cette figure illustre la comparaison des paramètres locaux du réseau réel *InteroFull* et de 50 instances de chacun des modèles aléatoires (*er*, *er_ dd*, *geo*, *sf*, *sticky*). Le graphe du haut montre la similarité entre les distributions des degrés des graphlets. Le graphe du bas montre la distance entre les fréquences relatives des graphlets.

utilisé une méthode classiquement utilisée pour identifier les modules d'un réseau d'interactions protéine-protéine, à savoir la méthode MCL. L'objectif était de déterminer les paramètres adaptés pour appliquer cette méthode à chaque réseau. C'est pourquoi nous avons mis en place une méthode de comparaison de différentes décompositions en modules, en utilisant les annotations fonctionnelles. Finalement, nous avons pu comparer les décompositions en modules pour les différents réseaux expérimentaux et prédits.

3.4.1 Méthodes d'extraction de modules

Dans le contexte de la recherche de modules fonctionnels, différentes méthodes de clustering ont été appliquées aux réseaux d'interactions de manière à détecter des sous-graphes fortement connectés. C'est le cas notamment des quatre méthodes suivantes :

- **MCL** Markov Clustering [Enright *et al.*, 2002]
- **RNSC** Restricted Neighborhood Search Clustering [King *et al.*, 2004]
- **SPC** Super Paramagnetic Clustering [Blatt *et al.*, 1996]
- **MCODE** Molecular Complex Detection [Bader et Hogue, 2003]

Brohée *et al.* ont fait une étude comparative des ces méthodes de clustering [Brohée et van Helden, 2006] en considérant notamment leur robustesse par rapport aux faux positifs et faux négatifs contenus dans les réseaux d'interactions protéine-protéine, ainsi que leur sensibilité aux paramètres choisis. Ils ont conclu que la méthode MCL était meilleure que les autres pour extraire les complexes des réseaux d'interactions protéine-protéine. Cette méthode est en particulier très robuste aux altérations du graphe. C'est pourquoi nous avons décidé d'utiliser la méthode MCL afin de comparer les complexes des réseaux étudiés.

MCL est un algorithme de clustering de graphes basé sur une approche non supervisée et sur la simulation de flots dans le graphe. Cet algorithme consiste en un processus itératif. À chaque itération, des étapes d'expansion et d'inflation sont réalisées à l'aide de la matrice d'adjacence associée au graphe. D'abord, l'expansion consiste à simuler le flot à travers le graphe. Ensuite, l'inflation a pour but d'augmenter le contraste entre les régions de flots forts et de flots faibles du graphe. À la fin du processus, une partition du graphe est obtenue avec, d'un côté, un ensemble de régions parcourues par des flots importants, et de l'autre côté, des frontières sans flots. Ceci permet de définir une partition du graphe qui met en évidence les régions denses.

Un paramètre dénommé *inflation* permet de moduler l'étape d'inflation. Il a une influence sur la granularité de la partition obtenue. Ce paramètre varie en général entre 1,2 et 5. Plus la valeur du paramètre est élevée, plus le nombre de classes obtenues est grand.

Il est de plus possible de limiter l'expansion, c'est-à-dire de favoriser la tendance des nœuds à s'attacher à eux-même plutôt qu'aux voisins, en augmentant le paramètre de *centering*.

Il a été montré que la méthode MCL est sensible entre autres au choix de ces deux paramètres d'*inflation* [Brohée et van Helden, 2006] et de *centering* [Hart *et al.*, 2007]. Par conséquent, nous avons décidé, dans un premier temps, d'explorer l'espace produit des paramètres (*inflation*, *centering*), afin d'extraire les complexes de chaque réseau de

manière optimale, c'est-à-dire en choisissant les paramètres de manière pertinente.

3.4.2 Détermination des paramètres optimaux

Afin de déterminer les paramètres optimaux pour chacun des réseaux, nous avons exploré l'espace suivant :

- $1,4 \leq Inflation \leq 5,0$, par incréments de 0,2
- $0,5 \leq Centering \leq 2,25$, par incréments de 0,25

Nous avons alors utilisé les annotations fonctionnelles des protéines, afin d'évaluer chaque décomposition en modules, et de les comparer quantitativement. Pour cela, nous avons calculé un indice de dissimilarité fonctionnelle BGD (voir Équation 3.5), d'autant plus faible que la décomposition est cohérente sur le plan fonctionnel.

$$BGD = \frac{\sum_i |C_i|(1 - S(C_i)) + |\overline{C}|(1 - S(\overline{C}))}{|C| + |\overline{C}|} \tag{3.5}$$

où

- C est l'ensemble des classes qui contiennent au moins deux protéines
- C_i est l'ensemble des protéines appartenant à la classe i
- $\overline{C}$ est l'ensemble des protéines qui sont dans une classe singleton
- $S(X)$ est la mesure de similarité de l'ensemble des protéines X

La mesure de similarité fonctionnelle d'un groupe de protéines a été définie comme la moyenne des similarités de toutes les paires de protéines de l'ensemble (voir Équation 3.6).

$$S(C_i) = \frac{2}{|C_i|(|C_i| - 1)} \sum_{p_k, p_l \in C_i} ss(p_k, p_l) \tag{3.6}$$

où $ss(p_k, p_l)$ est la similarité fonctionnelle entre deux protéines. Cette similarité est définie sur la base des ensembles de termes GO associés à chacune des deux protéines [Lubovac *et al.*, 2006] (voir Annexe E).

Dans la mesure où certaines protéines ne sont annotées par aucun terme GO, les calculs ont été réalisés sur les sous-ensembles de paires de protéines annotées par au moins un terme pour chacune des deux protéines. Lorsque cette annotation minimale n'était pas disponible pour une paire de protéines de la classe C_i, le score $S(C_i)$ était égal à la moyenne sur l'ensemble des scores S de la classes C_i.

Pour chaque décomposition en modules, nous avons calculé la moyenne des scores BGD obtenus sur la base des annotations des deux ontologies *Molecular Function* et *Biological Process*. Nous avons représenté ces ensembles de score par des cartes de chaleur (heatmap). Nous en donnons deux exemples pour les réseaux *InteroPorc* (voir Figure 3.13) et *SatoCore* (voir Figure 3.14). Les scores minimaux ont alors été choisis pour sélectionner les meilleurs paramètres pour le calcul des modules. Les paramètres choisis sont indiqués dans la table 3.4.

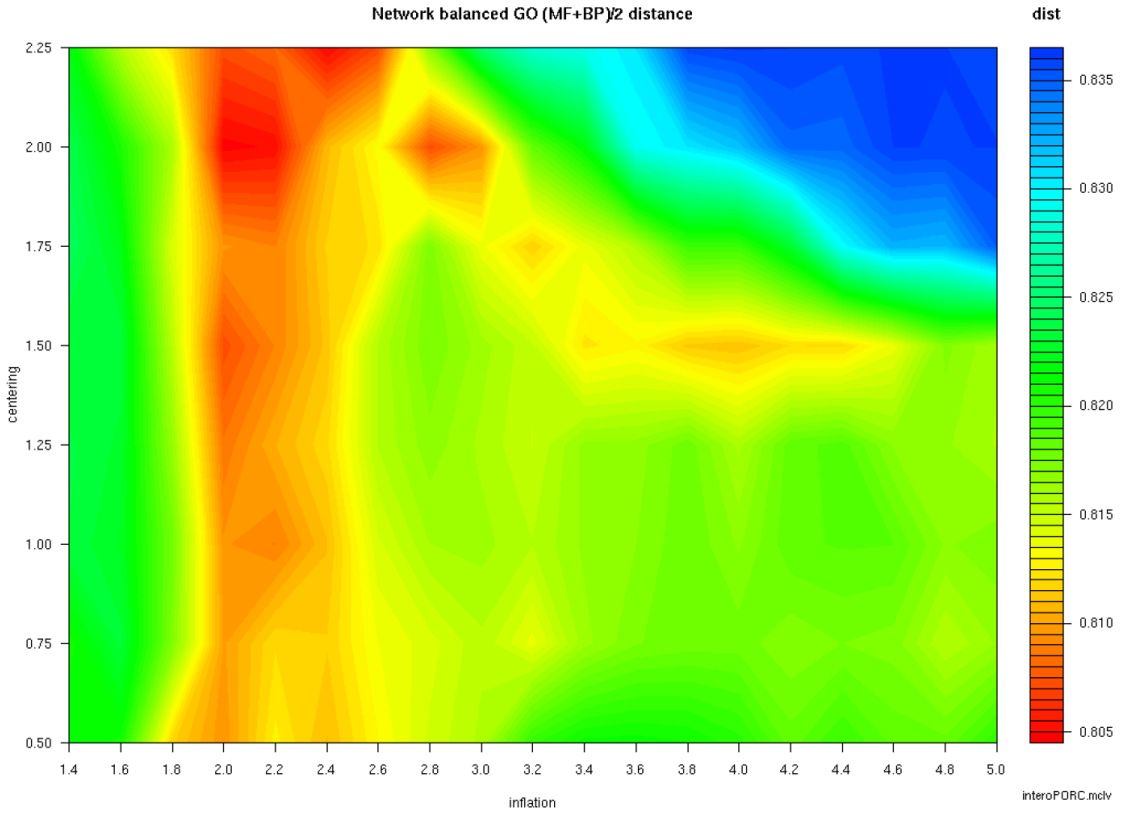


FIG. 3.13 – **Étude de l'influence des paramètres sur le réseau *InteroPorc*.** L'espace des paramètres est ici représenté avec l'*inflation* en abscisses et le *centering* en ordonnées. Pour chaque couple de valeurs, une décomposition en modules a été réalisée avec l'algorithme MCL. Cette décomposition a été évaluée par un score moyen basé sur les regroupements fonctionnels à partir des ontologies MF et BP de Gene Ontologie. Le score est représenté par la couleur. Il est d'autant plus faible que la décomposition est cohérente sur le plan fonctionnel.

Réseaux	**Inflation**	**Centering**
SatoFull	1,8	1
SatoCore	1,8	1,75
InteroFull	3,8	0,5
InteroPorc	2	2
InteroBH_LOW	3,8	0,5
InteroBH_MEDIUM	3,2	0,5
InteroBH_HIGH	3,2	1

TAB. 3.4 – **Choix des paramètres pour l'algorithme MCL.** Pour chaque réseau, nous avons choisi une valeur d'inflation et une valeur de centering. Ces valeurs ont été choisies de manière à optimiser la cohérence fonctionnelles des modules obtenus.

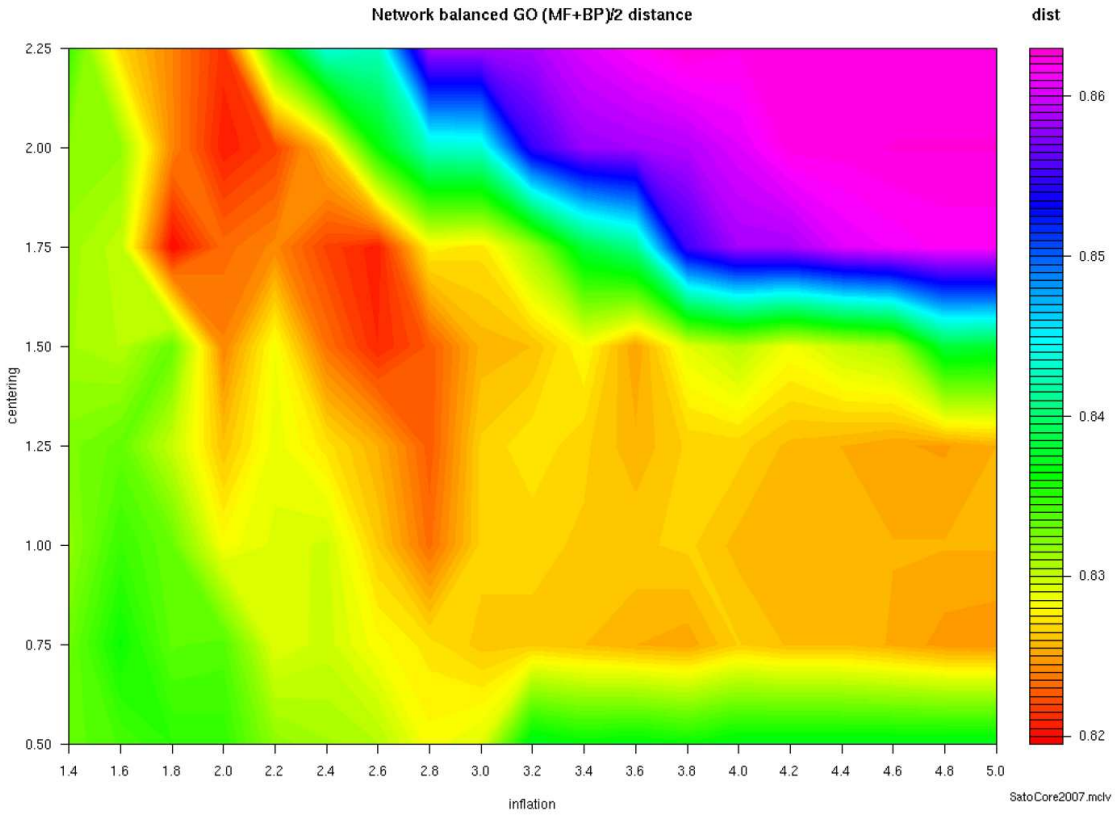


FIG. 3.14 – **Étude de l'influence des paramètres sur le réseau *SatoCore*.** L'espace des paramètres est ici représenté avec l'*inflation* en abscisses et le *centering* en ordonnées. Pour chaque couple de valeurs, une décomposition en modules a été réalisée avec l'algorithme MCL. Cette décomposition a été évaluée par un score moyen basé sur les regroupements fonctionnels à partir des ontologies MF et BP de Gene Ontologie. Le score est représenté par la couleur. Il est d'autant plus faible que la décomposition est cohérente sur le plan fonctionnel.

3.4.3 Comparaison des résultats

Après avoir déterminé les paramètres optimaux pour effectuer la recherche des modules adaptée à chaque réseau d'interactions protéine-protéine, nous avons pu réaliser ces décompositions et comparer les résultats obtenus (voir Table 3.5).

Nous avons noté que plus de modules ont été extraits des réseaux expérimentaux que des réseaux prédits (voir Table 3.5). Dans la mesure où *SatoCore* est beaucoup moins dense que *SatoFull* (voir Section 3.3), nous nous attendions à des résultats très différents pour ces deux graphes. Pourtant, des nombres très proches de modules ont été extraits pour ces deux graphes (455 et 555).

Les réseaux prédits mettent en évidence des tailles maximales largement supérieures à celles des réseaux expérimentaux. Ceci montre la présence d'au moins une grande classe dans laquelle se trouvent de nombreuses protéines, jusqu'à 287 pour le réseau *InteroBH_HIGH*. Par ailleurs, les tailles moyennes des modules extraits des réseaux prédits sont du même ordre de grandeur que celles des réseaux expérimentaux, à l'exception du réseau *InteroBH_HIGH* dont la taille moyenne s'élève à plus de 6. Ceci provient en partie de la classe géante citée précédemment et rassemblant 287 protéines d'un réseau qui en contient 741 (38,7%). Ceci peut s'expliquer par le fait que les réseaux prédits sont plus denses que les réseaux expérimentaux. Ceci peut également provenir des différences de paramètres pour les différentes décompositions.

Nous avons décomposé les réseaux d'interactions protéine-protéine expérimentaux et prédits en modules en utilisant l'algorithme MCL. Les résultats de ces décompositions montrent que les réseaux expérimentaux sont décomposés en modules de tailles assez homogènes, contenant entre deux et quatre protéines. Par contre, les réseaux prédits ont tendance à donner naissance à une grande classe contenant jusqu'à 39% des protéines du réseau.

Réseau	**Prot**	**PPI**	**C**	**Max**	**Moy**
SatoFull	1 920	3 236	555	41	3,46
SatoCore	1 152	1 064	455	9	2,53
InteroFull	1 011	8 783	401	137	2,52
InteroPorc	384	1 446	128	41	3,00
InteroBH_HIGH	741	2 748	122	287	6,07
InteroBH_MEDIUM	884	5 070	217	234	4,07
InteroBH_LOW	998	8 586	392	140	2,55

TAB. 3.5 – **Décomposition en modules des réseaux expérimentaux et prédits.** L'extraction des modules a été réalisée avec la méthode MCL [Enright *et al.*, 2002]. La colonne **Réseau** indique le nom de chaque réseau considéré. Les colonnes **Prot** et **PPI** indiquent respectivement les nombres de protéines et d'interactions contenues dans ces réseaux. La colonne **C** indique le nombre de modules obtenus après la décomposition par l'algorithme MCL. La colonne **Max** indique la taille du plus grand module (c'est-à-dire le nombre de protéines qui en font partie). La colonne **Moy** indique la taille moyenne des modules.

Conclusion

Au cours de ce chapitre, nous avons comparé les réseaux d'interactions protéine-protéine expérimentaux chez *Synechocystis* avec d'autres études expérimentales à grande échelle chez la levure et *H. pylori*. Nous avons pour cela proposé des méthodes d'analyse systématique de ces études expérimentales basées sur les rôles des protéines (appâts et proies). Ces méthodes permettent notamment d'analyser l'espace des interactions testées et les biais systématiques associées aux interactions ou aux protéines et provenant de difficultés techniques. Même si les méthodes proposées sont simples, elles ont l'avantage de prendre en compte de manière formelle des indicateurs techniques, ce qui ouvre des portes vers l'analyse automatique.

D'une part, l'analyse de la topologie des réseaux expérimentaux et prédits a mis en avant des points communs entre tous les réseaux, tels que la modularité et la correspondance avec le même modèle de graphe aléatoire (sticky). D'autre part, nous avons montré que les réseaux prédits sont malgré tout plus denses que les réseaux expérimentaux, conduisant à des décompositions en modules moins homogènes. Ce projet est actuellement encore en cours. Nous nous intéressons en particulier à l'analyse fonctionnelle des différents réseaux et une publication est en cours d'écriture [Michaut *et al.*, 2008a] (voir la liste des publications page 260).

À cette étape, nous avions un réseau global d'interactions protéine-protéine chez *Synechocystis* obtenu par des méthodes de prédiction *in-silico* et des méthodes d'identification expérientales. Ces différentes approches permettent d'obtenir des réseaux différents en termes de topologie et de modules. Néanmoins, ces cartes d'interactions protéine-protéine nous donnent une vision statique des interactions, alors que les interactions peuvent ne pas se produire en même temps, mais l'une à la la suite de l'autre, ou encore l'une à la place de l'autre. De plus, certaines interactions peuvent n'avoir lieu que dans certaines conditions, par exemple en réponse à un stress oxydant ou métallique.

L'idée a donc été de compléter ces cartes d'interactions avec les analyses des réponses transcriptionnelles obtenues chez *Synechocystis* en réponse à des stress oxydants et métalliques.

Chapitre 4

Étude de la dynamique des relations entre protéines

"L'ADN fonctionne d'une manière mystérieuse."

Richard Dawkins,
Le gène égoïste, 1976

L'objectif de ce chapitre était d'étudier la dynamique des relations entre les protéines. Nous avions en effet construit un réseau d'interactions protéine-protéine chez *Synechocystis* grâce à différentes méthodes de prédictions. Nous disposions également d'un réseau interactions protéine-protéine mises en évidence expérimentalement. Or, nous avions précédemment analysé la régulation de la transcription au cours du temps, en réponse à des stress oxydants et métalliques, notamment le cadmium et le peroxyde d'hydrogène. Ainsi, l'idée était de combiner cette information avec les réseaux d'interactions, afin d'ajouter une notion de dynamique, notamment en identifiant des modules fonctionnels dont les protéines étaient impliquées dans les réponses aux stress étudiés. Pour cela, nous avons d'abord voulu extraire des groupes de protéines en interaction et co-exprimées. Le but a ensuite été de caractériser la dynamique de ces groupes, c'est-à-dire de voir comment ils évoluaient au cours d'une régulation.

4.1 Identification de modules co-exprimés

L'idée était d'extraire des groupes de protéines en interaction et co-exprimées en réponse aux différents stress étudiés. Pour cela, nous nous sommes basés sur les différents réseaux d'interactions protéine-protéine présentés précédemment (voir Chapitres 2 et 3), obtenus par prédiction *in-silico* [Michaut *et al.*, 2008d] et par identification expérimentale [Sato *et al.*, 2007], ainsi que sur les données transcriptome étudiées précédemment (voir Chapitre 1). Nous avons ensuite voulu identifier, parmi ces différents modules, ceux qui étaient régulés, c'est-à-dire dont l'expression des gènes codant les protéines appartenant à ces modules était régulée au cours des différentes réponses transcriptionnelles. Enfin, nous avons centré notre analyse sur des protéines d'intérêt, notamment impliquées dans l'homéostasie du fer.

4.1.1 Extraction de modules

Dans le contexte de la recherche de modules fonctionnels, différentes méthodes de clustering ont été appliquées aux réseaux d'interactions de manière à détecter des sous-graphes fortement connectés. Nous avons choisi d'utiliser l'algorithme MCL [Enright *et al.*, 2002] qui est bien adapté à l'extraction de complexes des réseaux d'interactions protéine-protéine [Brohée et van Helden, 2006]. Pour chaque réseau, nous avons choisi les mêmes paramètres que précédemment (voir Table 3.4).

Pour prendre en compte la co-expression entre les protéines, l'idée a été de favoriser le rapprochement des protéines co-exprimées lors de la construction des modules. Pour cela, nous avons pondéré les interactions protéine-protéine avant d'effectuer l'extraction de modules par MCL, en nous basant sur les données transcriptome. En effet, nous avons quantifié la ressemblance des profils d'expression de deux gènes en calculant le coefficient de corrélation de Pearson (PCC). Ensuite, chaque interaction entre deux protéines a été pondérée par un score basé sur la ressemblance des profils d'expression des deux gènes codant les protéines en interaction (voir Équation 4.1).

$$S = -ln(\frac{1 - PCC}{2}), PCC \in]-1, 1[\tag{4.1}$$

Si le coefficient de corrélation de Pearson n'est pas défini à cause des valeurs manquantes ou s'il vaut 1 en valeur absolue, ce qui n'arrive pas dans la pratique, le score S n'est pas défini. L'algorithme MCL favorise la création de modules dont les interactions ont un poids élevé. Ainsi, les protéines en interaction avec un score S élevé ont tendance à être regroupées dans un même module, formant de cette façon des groupes de protéines en interaction et co-exprimées.

D'une part, nous avons étudié les réponses au Cd et à H_2O_2 en nous basant sur les cinétiques et leurs deux phases. D'autre part, nous avons étudié la réponse au Fe en regroupant les données en réponse à l'excès et à la carence en Fe. Nous n'avons pas pu considérer le zinc car le coefficient de corrélation de Pearson n'était pas pertinent dans ce cas par manque de points. Les résultats obtenus sont indiqués dans la table 4.1.

Nous avons tout d'abord remarqué que la prise en compte des données d'expression modifiait largement la décomposition en modules de chaque réseau d'interactions

PPI	**Stress**	**C**	**Max**	**Moy**	**Sd**
InteroFull	-	401	137	2,52	7,88
InteroPorc	-	128	41	3,00	4,52
SatoFull	-	555	41	3,46	2,96
SatoCore	-	455	9	2,53	0,97
InteroFull	Cd	425	48	2,38	4,11
InteroPorc	Cd	126	27	3,05	3,71
SatoFull	Cd	670	32	2,87	3,12
SatoCore	Cd	515	17	2,24	1,69
InteroFull	H_2O_2	395	63	2,56	4,23
InteroPorc	H_2O_2	126	33	3,05	4,15
SatoFull	H_2O_2	609	29	3,15	3,20
SatoCore	H_2O_2	471	14	2,45	1,83
InteroFull	Fe	379	104	2,67	6,34
InteroPorc	Fe	126	45	3,05	4,47
SatoFull	Fe	600	22	3,20	3,02
SatoCore	Fe	481	14	2,40	1,68

TAB. 4.1 – **Identification de modules.** Ces classes **C**, de taille moyenne **Moy** ont été identifiées à l'aide de l'algorithme MCL [Enright *et al.*, 2002]. Les interactions ont été pondérées pour chaque cinétique par la similarité des profils d'expression des gènes correspondant aux deux protéines en interaction. La colonne **PPI** indique le réseau d'interactions protéine-protéine considéré. La colonne **Stress** indique la cinétique ayant servi à pondérer les interactions. La première partie n'a pas été pondérée, il s'agit des décompositions obtenues de manière statique au Chapitre 3. La colonne **Max** indique la taille de la classe la plus grande, et **Sd** l'écart-type de la distribution des tailles des modules.

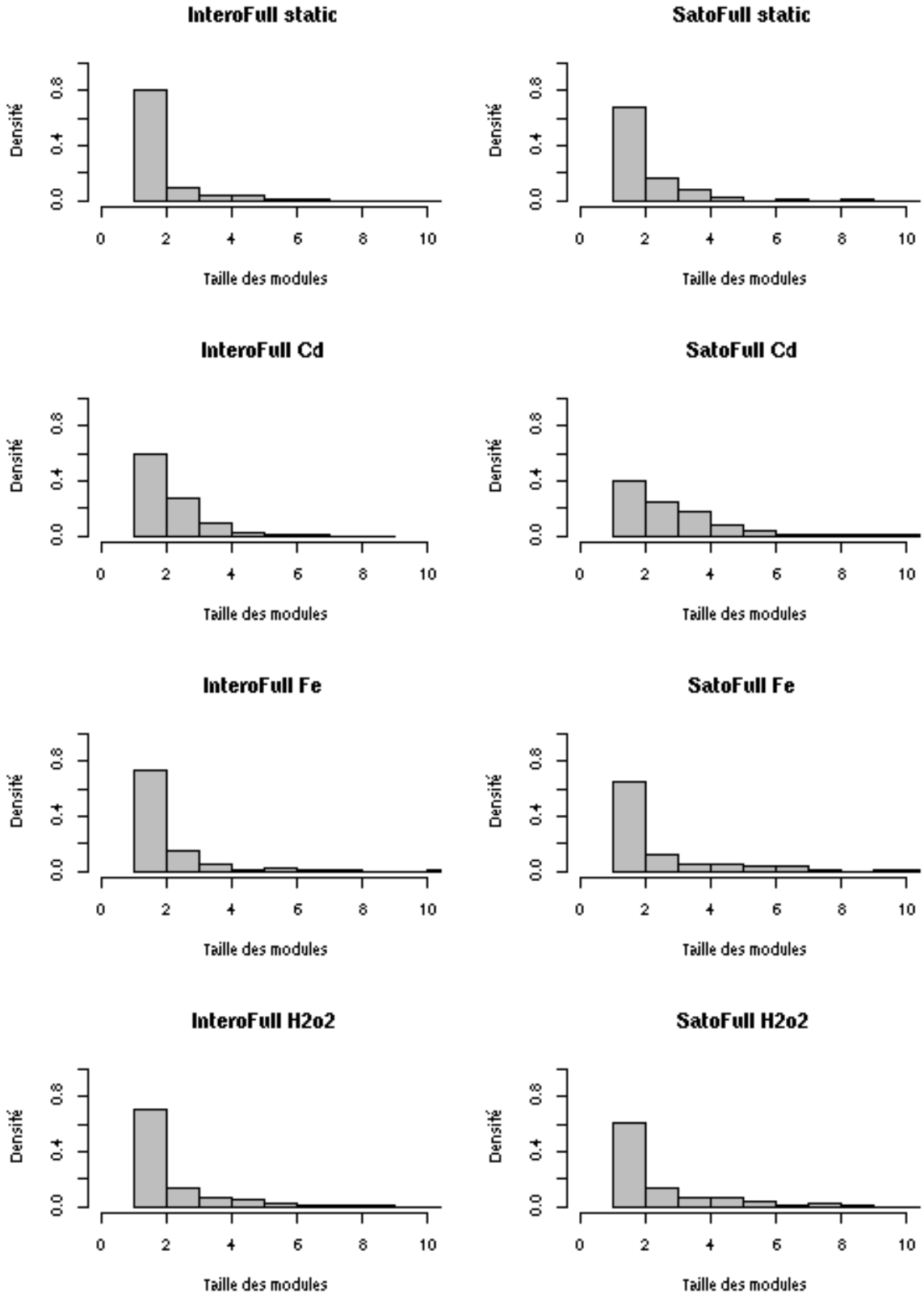


FIG. 4.1 – **Distributions des tailles des modules.** Ce graphe représente les distributions tronquées des tailles des modules pour différentes décompositions des réseaux *InteroFull* et *SatoFull*, de manière statique (les deux du haut), ou avec la pondération des données transcriptome (les six en-dessous).

protéine-protéine par rapport à celle effectuée précédemment, basée uniquement sur les interactions protéine-protéine (voir Table 4.1). Dans le cas des réseaux prédits, la décomposition obtenue avec pondération est constituée d'un peu plus de modules dont les tailles sont plus homogènes. En effet, la taille maximale est largement plus faible dans cette nouvelle décomposition, réduisant ainsi la disparité (voir Table 4.1 et Figure 4.1). L'absence de grande composante, comme dans les décompositions sans pondération pour le réseau *InteroFull*, peut s'expliquer par le fait que ces grands groupes sont scindés en plusieurs en fonction du type de régulation des différentes protéines. En revanche, les nouvelles décompositions des réseaux expérimentaux présentent des modules plus grands dont les tailles sont moins homogènes (voir Table 4.1 et Figure 4.1).

D'autre part, nous avons remarqué que les décompositions obtenues en se basant sur les données d'expression Cd et H_2O_2 étaient relativement proches. Ceci n'est pas très étonnant dans la mesure où ces deux conditions induisent une régulation de beaucoup de gènes en commun.

Ainsi, cette méthode permet de prendre en compte les données transcriptome et les données d'interaction afin de regrouper des protéines en modules fonctionnels.

4.1.2 Extraction de modules régulés

L'idée a ensuite été d'extraire parmi les modules ceux qui étaient régulés au cours des différentes réponses transcriptionnelles. Pour cela, nous avons calculé les moyennes des log-ratio des différents gènes sur les deux phases de réponse (ou les deux conditions dans le cas du Fe). Nous avons alors sélectionné les modules dont le log-ratio moyen était supérieur à 0,8 (induit) ou inférieur à -0,8 (réprimé) au cours d'une des deux phases au moins. De cette façon, nous avons sélectionné à la fois les modules étant régulés au cours d'une phase, et ceux étant régulés tout au long de la cinétique. Les résultats obtenus sont indiqués dans la table 4.2.

Il est intéressant de noter qu'il y a environ deux fois plus de modules régulés en pourcentage pour la cinétique H_2O_2 par rapport à la cinétique Cd, et ceci pour chaque réseau d'interactions protéine-protéine. Ceci provient probablement du fait que la cinétique Cd est plus longue que la cinétique H_2O_2. Par conséquent, la moyenne des log-ratios est plus faible, et moins de modules ont une valeur moyenne au-dessus du seuil fixé. Pour cette raison, il conviendrait de déterminer de manière plus fine la sélection des modules régulés.

Dans le cas du Fe, beaucoup moins de gènes sont régulés. C'est pourquoi peu de modules sont également régulés (voir Tables 1.2 et 1.3).

4.1.3 Extraction de modules impliqués dans l'homéostasie du fer

Nous nous sommes finalement intéressés aux protéines impliquées dans des mécanismes d'intérêt pour le laboratoire, en particulier l'homéostasie du fer. En effet, l'assemblage des protéines à centre fer-soufre est très contrôlé et met en œuvre une machinerie protéique complexe. Ces protéines à centres fer-soufre renferment des groupes très structurés d'atomes de fer et de soufre. Ces composés sont des transporteurs à électrons.

Pour étudier ces processus d'intérêt, une liste de protéines a été établie, contenant 129 protéines (voir la Section I.2 de l'Annexe I). Nous avons alors sélectionné les modules contenant au moins une protéine de cette liste. Les résultats obtenus sont indiqués dans la table 4.3.

Nous avons alors voulu représenter ces informations de manière visuelle. Pour cela nous avons utilisé le logiciel Cytoscape [Shannon *et al.*, 2003] permettant la visualisation des réseaux biologiques. Nous avons ainsi pu représenter différentes informations telles que les interactions protéine-protéine, les modules, la régulation de l'expression de chaque module, ainsi que la description des différentes protéines (voir Figures 4.2 et 4.3).

Nous avons illustré les résultats en réponse à la carence et à l'excès de Fe sur les réseaux prédits et expérimentaux (voir Figures 4.4 et 4.5).

Dans le cas du réseau *InteroFull*, en réponse à l'excès et la carence en Fe (voir Figure 4.4), trois modules ont été mis en évidence, contenant deux, trois et quatre protéines. Le premier module (en bleu) contient une protéine à centre fer-soufre codée par petF (*ssl0020*) qui est en interaction avec une protéine non caractérisée jusqu'à présent dans la base de données Uniprot (*slr1300*). La première est une ferredoxine qui a un rôle de médiateur dans le transfert des électrons. Nous pouvons alors émettre l'hypothèse que la seconde protéine joue un rôle dans le processus du transfert des électrons. Ceci semble tout à fait pertinent puisque la base de données Cyanobase reporte une forte homologie avec une protéine rédox (*similar to 2-octaprenyl-6-methoxyphenol hydroxylase*).

Le deuxième module (en rose) contient également la protéine sodB qui lie du Fe (*slr1516*). En effet, le Fe catalyse la réaction de dismutation, et permet à la superoxide dismutase de détruire les radicaux toxiques qui sont produits dans la cellule. Il est intéressant de noter que ce module est induit en cas d'excès de Fe, et réprimé en cas de carence en Fe. Ainsi, l'expression du gène *slr1516* est en phase avec la disponibilité en Fe. Cela dit, il est étonnant que la protéase *slr0535* semble co-exprimée avec la superoxide dismutase.

Le troisième module (en vert) contient quant à lui un transporteur membranaire (*sll1768*). Lorsqu'il y a trop de fer, ce module est réprimé, sans doute pour limiter les entrées en Fe dans la cellule, alors qu'il est au contraire induit en cas de carence en Fe, sans doute pour favoriser les entrées de Fe. Les deux protéines inconnues (*slr0903* et *sll1524*) sont probablement impliquées également dans le transport du Fe. Il serait intéressant d'explorer ces pistes plus en détail.

Par ailleurs, il est intéressant de noter que toutes les protéines (36) du sous-graphe du réseau *InteroFull* sont caractérisées, alors que 16 protéines sur les 47 du sous-graphe du réseau *SatoFull* sont inconnues. De plus, l'analyse de récentes analyses transcriptomiques en réponse au manque de soufre pourrait permettre d'approfondir cette étude [Zhang *et al.*, 2008].

PPI	Stress	C	%	Moy	Max
InteroFull	Cd	29	6,8	6	2,69
InteroPorc	Cd	6	4,7	5	2,83
SatoFull	Cd	41	6,1	9	3,22
SatoCore	Cd	45	8,7	6	2,73
InteroFull	H_2O_2	70	17,7	12	2,96
InteroPorc	H_2O_2	15	11,9	10	3,67
SatoFull	H_2O_2	84	13,8	12	3,40
SatoCore	H_2O_2	69	14,6	7	2,71
InteroFull	Fe	24	6,3	9	3,25
InteroPorc	Fe	4	3,2	6	3,50
SatoFull	Fe	14	2,3	7	3,14
SatoCore	Fe	16	3,3	6	2,81

TAB. 4.2 – **Identification de modules régulés.** Parmi les complexes identifiés (voir Table 4.1), nous avons extrait les modules régulés, de taille moyenne **Moy**, en considérant la valeur moyenne sur l'ensemble des protéines de chaque module du log-ratio moyen au cours de la cinétique considérée. Comme précédemment, la colonne **PPI** indique le réseau d'interactions protéine-protéine considéré. La colonne **Stress** indique la cinétique ayant servi à pondérer les interactions. Les colonnes **C** et **%** indiquent respectivement le nombre de modules régulés et le pourcentage qu'ils représentent par rapport à l'ensemble des modules identifiés. La dernière colonne **Max** indique la taille du module le plus grand.

PPI	Stress	C	Moy	Max
InteroFull	Cd	3	2	2,00
InteroPorc	Cd	0	0	0,00
SatoFull	Cd	6	5	3,17
SatoCore	Cd	2	3	2,50
InteroFull	H_2O_2	12	7	3,00
InteroPorc	H_2O_2	0	0	0,00
SatoFull	H_2O_2	9	12	5,22
SatoCore	H_2O_2	2	7	4,50
InteroFull	Fe	3	4	3,00
InteroPorc	Fe	1	2	2,00
SatoFull	Fe	4	4	2,75
SatoCore	Fe	2	2	2,00

TAB. 4.3 – **Identification de modules régulés avec protéines d'intérêt.** Parmi les modules régulés (voir Table 4.2), nous avons extrait les modules impliqués notamment dans l'homéostasie du fer, en considérant une liste de protéines d'intérêt. Comme précédemment, la colonne **PPI** indique le réseau d'interactions protéine-protéine considéré. La colonne **Stress** indique la cinétique ayant servi à pondérer les interactions. La colonne **C** indique le nombre de modules régulés. Leur taille moyenne est donnée dans la colonne **Moy**. La dernière colonne **Max** indique la taille du module le plus grand.

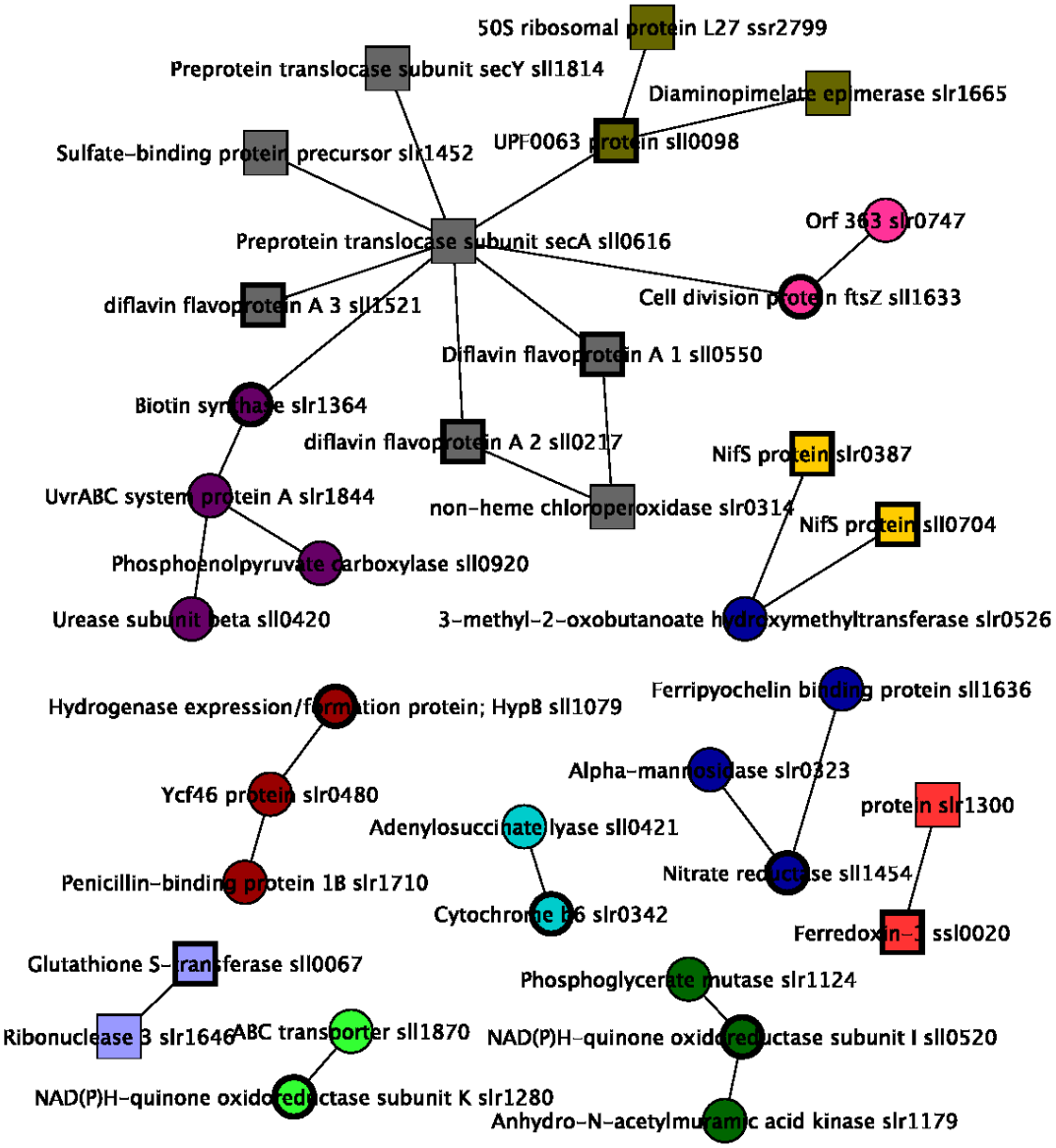


FIG. 4.2 – **Visualisation des modules du réseau *InteroFull*.** Ce graphique a été réalisé à l'aide du logiciel Cytoscape [Shannon *et al.*, 2003] ; il représente les 12 modules régulés du réseau *InteroFull* et contenant des protéines d'intérêt dont la bordure est en gras. Chaque module est caractérisé par une couleur. Les carrés représentent les protéines codées par des gènes induits, et les cercles les protéines codées par des gènes réprimés, au cours de la cinétique H_2O_2. La description de chaque protéine est inscrite sur le nœud la représentant. Le sous-réseau contient 28 interactions entre 36 protéines.

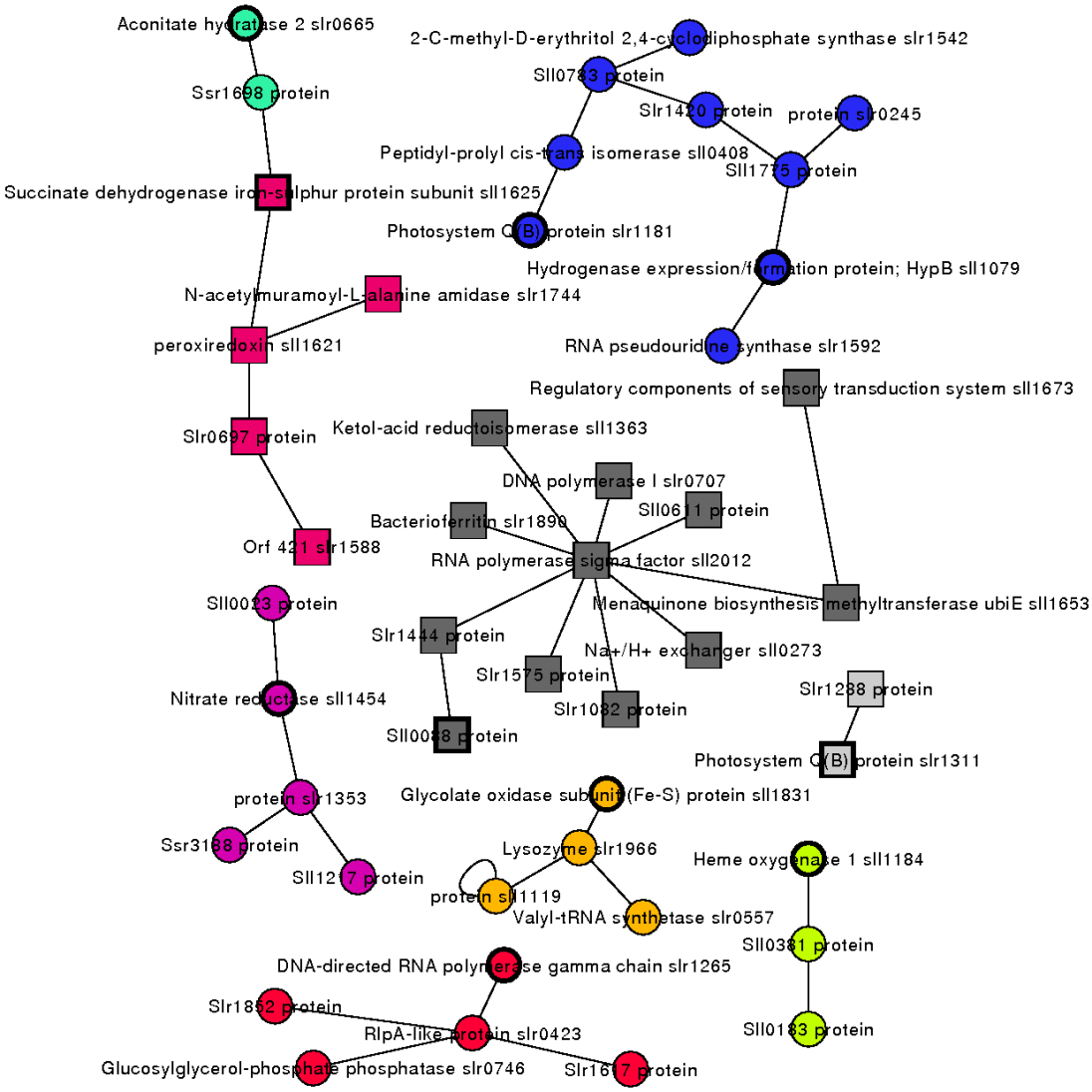


FIG. 4.3 – **Visualisation des modules du réseau *SatoFull*.** Ce graphique a été réalisé à l'aide du logiciel Cytoscape [Shannon *et al.*, 2003] ; il représente les neuf modules régulés du réseau *SatoFull* et contenant des protéines d'intérêt dont la bordure est en gras. Chaque module est caractérisé par une couleur. Les carrés représentent les protéines codées par des gènes induits, et les cercles les protéines codées par des gènes réprimés, au cours de la cinétique H_2O_2. La description de chaque protéine est inscrite sur le nœud la représentant. Le sous-réseau contient 40 interactions entre 47 protéines.

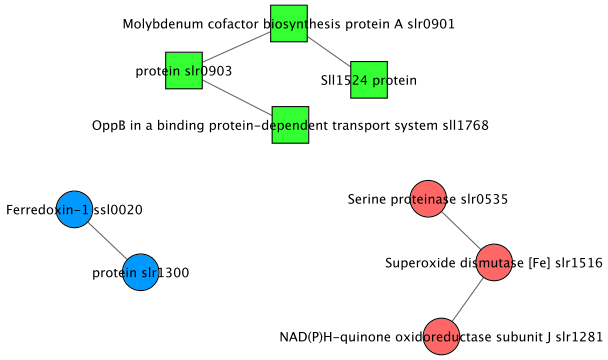


FIG. 4.4 – **Visualisation des modules du réseau *InteroFull*.** Ce graphique a été réalisé à l'aide du logiciel Cytoscape [Shannon *et al.*, 2003] ; il représente les trois modules régulés du réseau *InteroFull* et contenant des protéines d'intérêt. Chaque module est caractérisé par une couleur. Les carrés représentent les protéines codées par des gènes induits, et les cercles les protéines codées par des gènes réprimés, au cours de la cinétique Fe. La description de chaque protéine est inscrite sur le nœud la représentant. Le sous-réseau contient six interactions entre neuf protéines.

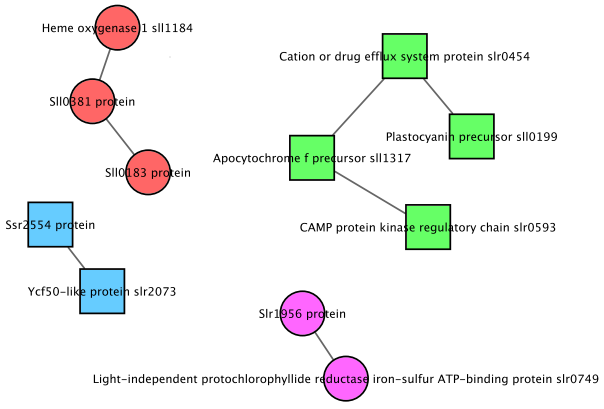


FIG. 4.5 – **Visualisation des modules du réseau *SatoFull*.** Ce graphique a été réalisé à l'aide du logiciel Cytoscape [Shannon *et al.*, 2003] et représente les trois modules régulés du réseau *SatoFull* et contenant des protéines d'intérêt. Chaque module est caractérisé par une couleur. Les carrés représentent les protéines codées par des gènes induits, et les cercles les protéines codées par des gènes réprimés, au cours de la cinétique Fe. La description de chaque protéine est inscrite sur le nœud la représentant. Le sous-réseau contient sept interactions entre 11 protéines.

Nous avons mis en place une méthode permettant de décomposer un réseau d'interactions protéine-protéine en modules, et qui présente l'avantage de prendre en compte la régulation de l'expression des gènes en réponse à différents stress. De plus, nous proposons ici une visualisation des résultats de manière globale. Ceci permet d'explorer certains sous-réseaux d'intérêts, et de mettre en évidence des protéines inconnues impliquées par exemple dans l'homéostasie du fer et du soufre.
Néanmoins, les informations contenues dans les cinétiques sont ici réduites en une similarité de profils. Par conséquent, une partie de l'information est perdue et nous ne pouvons pas étudier la dynamique de ces différents modules au cours des cinétiques. Nous avons alors voulu développer une autre méthode palliant ce manque.

4.2 Caractérisation de la dynamique des modules

Nous avons voulu caractériser par la suite la dynamique des modules au cours des cinétiques, par exemple voir si des modules sont régulés à certains moments de la cinétique, si des modules se créent ou disparaissent. Pour cela, nous avons construit une décomposition en modules propre à chaque temps d'une cinétique. Ensuite, nous avons analysé les transitions entre les temps de la cinétique, notamment en identifiant les différents événements ayant lieu entre les modules. Enfin, nous avons voulu identifier les modules constants au cours de la cinétique en termes de composition, ainsi que les composantes isolées entre les différents modules.

4.2.1 Création des modules

L'objectif était de caractériser les modules au cours d'une cinétique donnée. Pour cela, pour chaque temps de la cinétique, nous avons calculé une similarité entre les gènes en nous basant sur les log-ratios d'expression (voir Équation 4.2). Cette similarité a été utilisée pour pondérer les interactions protéine-protéine de chaque réseau considéré, afin de le décomposer en modules grâce à l'algorithme MCL, en tenant compte de l'expression des gènes correspondants.

$$S = \frac{1}{|a - b|} \tag{4.2}$$

où a et b sont les log-ratios quantifiant l'expression des gènes correspondant aux deux protéines en interaction.

Les résultats indiqués dans la table 4.4 caractérisent les décompositions en modules du réseau d'interactions protéine-protéine *InteroFull*, en réponse au stress Cd (voir les détails des autres réseaux et des autres stress dans l'annexe I).

En comparaison avec la décomposition précédente, où la pondération était faite de manière globale grâce à la similarité des profils (voir Table 4.1), les différentes décompositions présentent des modules légèrement plus grands en moyenne, et une taille maximum également plus grande. Ceci reste globalement valable pour les différents réseaux et les

différents stress. La différence majeure concerne la taille moyenne des modules pour le réseau *SatoCore* qui reste sensiblement la même que précédemment, et ceci pour les deux stress étudiés.

L'évolution de ces paramètres reste assez faible au cours de la cinétique. La taille moyenne des modules est globalement constante. La taille maximale varie un peu tout en restant du même ordre de grandeur. Aucun module aussi grand que ceux obtenus sans la pondération n'apparaît pour le réseau *InteroFull* (voir Table 4.1). Par contre, le réseau *SatoCore* possède de plus grands modules que ceux obtenus sans pondération. Ainsi, la pondération permet d'avoir des décompositions en modules plus homogènes entre les réseaux.

4.2.2 Analyse des transitions

Nous avons alors voulu analyser les transitions entre les différents temps de chaque cinétique, c'est-à-dire voir dans quelle mesure les modules étaient constants ou variaient. Pour cela, nous avons défini une relation de similarité entre deux modules de deux temps d'une cinétique (voir Définition 4.1).

Définition 4.1 (Relation de similarité entre modules) *Soit N un réseau d'interactions protéine-protéine. Soient T_1 et T_2 les deux partitions du réseau N obtenues pour deux temps d'une cinétique. Soit M_1 (resp. M_2), l'ensemble des modules de T_1 (resp. T_2). La relation de similarité entre deux modules est définie de la manière suivante :*

$$(\forall\, a \in M_1)\,(\forall\, b \in M_2)\,\{(a \mathcal{S} b) \iff (\forall\, m \in M_2)(h_{am} < h_{ab})\}$$

où h_{ij} est l'homologie entre les deux modules i et j. Elle est définie comme le nombre de protéines en commun entre les deux modules $h_{ij} = |i \cap j|$.

Notons que cette relation n'est pas symétrique. En faisant un parallèle avec la relation de similarité de séquences entre les protéines de deux espèces données, nous avons défini une relation symétrique en considérant la relation de similarité d'un module vers un autre et réciproquement.

Définition 4.2 (Relation de conservation entre modules) *Soit N un réseau d'interactions protéine-protéine. Soient T_1 et T_2 les deux partitions du réseau N obtenues pour deux temps d'une cinétique. Soit M_1 (resp. M_2), l'ensemble des modules de T_1 (resp. T_2). La relation de conservation entre deux modules est définie de la manière suivante :*

$$(\forall\, a \in M_1)\,(\forall\, b \in M_2)\,\{(a \Re b) \iff (a \mathcal{S} b \,\&\, b \mathcal{S} a)\}$$

Ainsi, nous avons utilisé cette relation pour définir la conservation de certains modules d'un temps à un autre d'une cinétique. Cet événement est dénoté P (pour *preserve*).

Nous avons alors défini quatre autres événements pour caractériser les transitions entre deux temps d'une cinétique :

M *merge* : deux modules sont regroupés
S *split* : un module se divise en deux
A *appear* : un module apparaît
D *disappear* : un module disparaît

Les événements identifiés dans le réseau *InteroFull*, pour chaque transition de la cinétique Cd, sont indiqués dans la table 4.5 (voir les détails des autres réseaux et des autres stress dans l'annexe I).

Tous les modules sont assignés à un événement exactement. Ainsi, au cours d'une transition entre deux partitions successives T_1 et T_2 du réseau, il existe des relations entre les tailles de ces deux partitions, c'est-à-dire le nombre de modules qu'elles contiennent, et les nombres d'événements. Ces relations sont décrites par les deux équations suivantes :

$$|T_1| = P + 2M + S + D$$
$$|T_2| = P + M + 2S + A$$

Nous pouvons alors représenter les transitions entre les modules sous la forme d'un graphe orienté de manière intuitive. La figure 4.6 montre le réseau *InteroFull* et l'évolution des modules au cours de la cinétique H_2O_2. Nous avons filtré les modules de taille inférieure à cinq. De plus, nous avons utilisé la couleur des nœuds du graphes pour représenter la régulation globale de chacun des modules. Ainsi, pour un module donné, nous avons considéré la valeur moyenne des log-ratios des gènes correspondants aux protéines composant ce module. Si cette valeur est supérieure à 0,1, le nœud est rouge ; si cette valeur est inférieure à -0,1, le nœud est vert ; sinon il est gris.

La visualisation de ce graphe de transitions permet d'avoir une image globale de l'évolution des modules au cours de la cinétique. Nous avons alors voulu savoir si certains modules étaient conservés tout au long de la cinétique.

4.2.3 Identification de modules stables

Le but était d'identifier les modules préservés au cours de chaque transition. Ces modules peuvent être considérés comme constants dans le temps ou stables. Néanmoins, même si les modules sont préservés au cours de la cinétique, leur composition peut varier un peu.

Davantage de modules stables ont été détectés au cours de la cinétique Cd qu'au cours de la cinétique H_2O_2 (voir Table 4.6). Toutefois, plus de modules stables de taille supérieure à 5 ont été identifiés au cours de la cinétique H_2O_2.

Les autres modules qui ne sont pas constants au cours des cinétiques se mélangent (*split*, *merge*), apparaissent et disparaissent. Nous avons alors voulu voir dans quelle mesure certains groupes de modules étaient isolés les uns des autres.

Réseau	**Stress**	**Tps**	**Prot**	**PPI**	**C**	**Max**	**Moy**
InteroFull	Cd	1	1 011	8 783	306	37	3,30
InteroFull	Cd	2	1 011	8 783	304	35	3,33
InteroFull	Cd	3	1 011	8 783	267	49	3,79
InteroFull	Cd	4	1 011	8 783	297	33	3,40
InteroFull	Cd	5	1 011	8 783	302	65	3,35
InteroFull	Cd	6	1 011	8 783	310	40	3,26
InteroFull	Cd	7	1 011	8 783	304	41	3,33
InteroFull	Cd	8	1 011	8 783	301	33	3,36
InteroFull	Cd	9	1 011	8 783	297	34	3,40

TAB. 4.4 – **Décomposition en modules par MCL pour chaque temps d'une cinétique.** Ces classes **C**, de taille moyenne **Moy**, ont été identifiées à l'aide de l'algorithme MCL [Enright *et al.*, 2002]. Les interactions ont été pondérées pour chaque temps **Tps** de la cinétique **Stress** par la similarité des taux d'expression des gènes correspondants aux deux protéines en interaction. La colonne **Réseau** indique le réseau d'interactions protéine-protéine considéré. Les colonnes **Prot** et **PPI** indiquent respectivement le nombre de protéines et d'interactions du réseau considéré. La colonne **Max** indique la taille de la classe la plus grande. L'ensemble des résultats est présenté en Annexe (voir Table I.1).

Transition	**T1**	**T2**	**Events**	**P**	**M**	**S**	**A**	**D**
1-2	306	304	347	151	34	22	65	75
2-3	304	267	332	156	28	15	80	56
3-4	267	297	337	155	20	16	56	90
4-5	297	302	333	174	25	21	52	61
5-6	302	310	346	176	30	15	51	74
6-7	310	304	331	179	35	17	44	56
7-8	304	301	334	179	27	19	52	57
8-9	301	297	339	153	35	18	60	73

TAB. 4.5 – **Évolution des modules d'un temps à l'autre.** Pour chaque transition d'un temps caractérisé par **T1** modules au suivant, caractérisé par **T2** modules, le nombre d'événements identifiés est indiqué dans la colonne **Events**. Nous indiquons ensuite les nombre d'événements identifiés dans chacune des catégories suivantes : conservation **P**, regroupement **M**, division **S**, apparition **A**, disparition **D**. L'ensemble des résultats est présenté en Annexe (voir Table I.3).

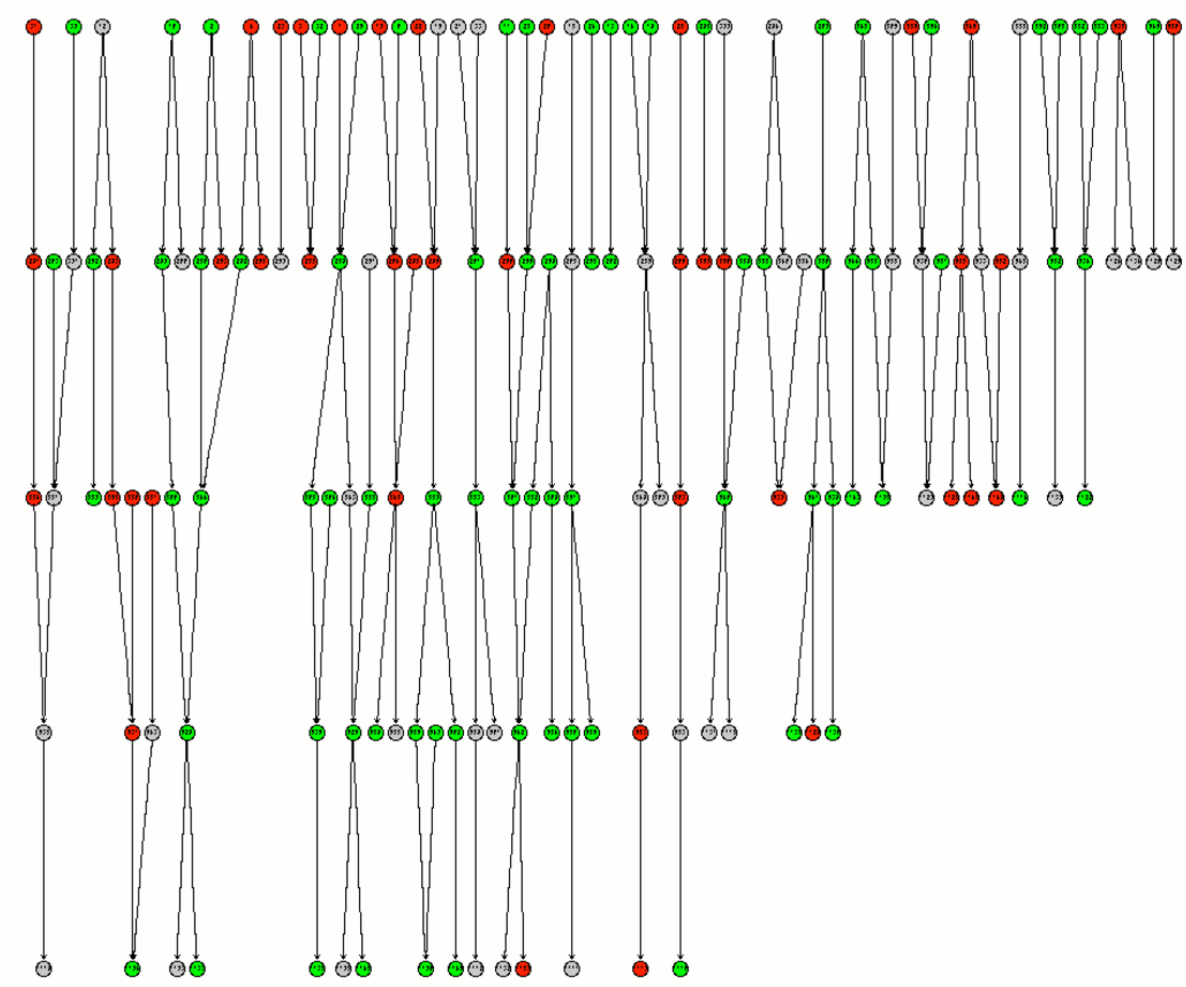


FIG. 4.6 – **Visualisation de l'évolution des modules.** Ce graphe représente les transitions entre les modules extraits du réseau *InteroFull*. Chaque nœud représente un module, tandis que les arêtes indiquent les différents événements identifiés au cours des transitions. Chaque niveau horizontal représente ainsi la décomposition en modules d'un temps de la cinétique H_2O_2. La couleur des modules met en évidence la régulation du taux d'expression moyen des gènes codant les différentes protéines du module (rouge=induit, vert=réprimé, gris=non-régulé).

4.2.4 Identification de groupes isolées

L'objectif était de séparer du graphe des transitions les composantes connexes, afin d'extraire les groupes de modules isolés les uns des autres. Pour cela, nous avons construit des groupes de modules, ou méta-modules, regroupant tous les modules reliés par des événements.

Les résultats montrent que les réseaux *SatoCore* et surtout *SatoFull* sont composés de nettement plus de composantes isolées que le réseau *InteroFull* (voir Table 4.7).

Le formalisme que nous proposons permet donc de dégager les groupes stables. En effet, chaque temps d'une cinétique constitue une légère perturbation par rapport à la décomposition globale du réseau. Les transitions mettent en évidence les changements. De plus, l'identification des groupes isolés permet de faire un consensus entre les différentes décompositions en modules.

Nous avons exploité les aspects dynamiques des cinétiques pour moduler les décompositions en modules des différents réseaux d'interactions protéine-protéine. Ceci nous a permis de caractériser l'évolution des modules, notamment en définissant des événements de transition entre les modules d'un temps d'une cinétique au suivant. De plus, nous avons identifié les modules stables au cours des cinétiques. Enfin, nous avons extrait des groupes de modules isolés.

Réseau	**Stress**	**Filtrage**	**Nb C**
InteroFull	Cd	0	100
SatoFull	Cd	0	343
SatoCore	Cd	0	347
InteroFull	Cd	5	0
SatoFull	Cd	5	8
SatoCore	Cd	5	13
InteroFull	H_2O_2	0	86
SatoFull	H_2O_2	0	277
SatoCore	H_2O_2	0	320
InteroFull	H_2O_2	5	4
SatoFull	H_2O_2	5	17
SatoCore	H_2O_2	5	25

TAB. 4.6 – **Identification de modules stables.** La colonne **Réseau** indique le réseau d'interactions protéine-protéine considéré. La colonne **Stress** indique la cinétique utilisée pour pondérer les interactions. La colonne **Filtrage** indique la taille en-dessous de laquelle les modules n'ont pas été pris en compte. La colonne **Nb C** indique le nombre de modules stables au cours de la cinétique.

Réseau	**Stress**	**Filtrage**	**Nb C**
InteroFull	Cd	0	347
SatoFull	Cd	0	637
SatoCore	Cd	0	493
InteroFull	Cd	5	92
SatoFull	Cd	5	184
SatoCore	Cd	5	92
InteroFull	H_2O_2	0	319
SatoFull	H_2O_2	0	576
SatoCore	H_2O_2	0	450
InteroFull	H_2O_2	5	83
SatoFull	H_2O_2	5	180
SatoCore	H_2O_2	5	96

TAB. 4.7 – **Identification de groupes isolés.** La colonne **Réseau** indique le réseau d'interactions protéine-protéine considéré. La colonne **Stress** indique la cinétique utilisée pour pondérer les interactions. La colonne **Filtrage** indique la taille en-dessous de laquelle les modules n'ont pas été pris en compte. La colonne **Nb C** indique le nombre de groupes isolés au cours de la cinétique, c'est-à-dire qui n'interagissent pas avec d'autres modules.

Conclusion

Dans ce chapitre, nous avons proposé des méthodes d'analyse des modules de protéines en nous basant sur des réseaux d'interactions protéine-protéine et des données d'expression des gènes correspondants. Nous avons ainsi identifié des modules pour les différents réseaux et les différents stress étudiés. Nous avons extrait les modules régulés et ceux faisant intervenir des protéines impliquées dans certains mécanismes d'intérêt, comme l'homéostasie du fer et du soufre. Ceci nous a permis de mettre en évidence des protéines inconnues potentiellement impliquées dans ces processus. De plus, nous avons proposé un formalisme pour étudier la dynamique de ces modules, et notamment les transitions entre les différents temps d'une cinétique.

Si ces méthodes sont encore très simples et très exploratoires, elles permettent néanmoins de combiner les données à l'échelle du génome pour se recentrer sur certains processus biologiques particuliers. Il conviendrait d'analyser en détails les résultats obtenus, notamment en termes fonctionnels, afin de les évaluer globalement et d'obtenir de nouvelles pistes en termes de biologie. Les méthodes proposées pourraient éventuellement être adaptées en fonction des observations fonctionnelles. D'autres méthodes de décomposition en modules pourraient également être considérées [Hwang *et al.*, 2006], [Ulitsky et Shamir, 2007].

D'autre part, il serait intéressant d'exploiter les réseaux d'interactions protéine-protéine non plus pour extraire des modules, mais plutôt des voies de signalisation [Segal *et al.*, 2003], [Baudot *et al.*, 2008].

Discussion

"La science n'est pas "le bon sens systématisé". Pour être passionnante, elle doit remettre en question notre vision du monde et opposer des théories solides aux vieux préjugés anthropocentristes que nous appelons intuition."

Stephen Jay Gould,
Darwin et les grandes énigmes de la vie, 1997

Le concept de classification a été utilisé à plusieurs reprises dans ce travail, que ce soit la classification mixte hiérarchique pyramidale, ou la classification fonctionnelle utilisée pour annoter les protéines.

Une classification est un système organisé et hiérarchisé, qui permet de classer les connaissances dans un domaine particulier. Les classifications sont utilisées entre autres dans les sciences de la nature (classification phylogénétique, tableau périodique des éléments) ou les bibliothèques (classification bibliothéco-bibliographique BBK créée du temps de l'Union Soviétique, principes de classement des documents musicaux PCDM, classification décimale de Dewey CDD). Certaines classifications, comme par exemple la classification hiérarchique, permettent de représenter des données notamment en hiérarchisant les relations entre les objets considérés. Ainsi, la structure hiérarchique est une suite d'inclusions qui décrit une organisation en différents niveaux. À chaque niveau, la classification hiérarchique définit une partition des objets.

Néanmoins, cette organisation hiérarchique suppose que ces différentes classes sont disjointes, c'est-à-dire qu'un élément ne peut appartenir qu'à une seule classe. Or, si l'on considère les protéines et leur implication dans une fonction donnée, on est confronté à un problème. En effet, la plupart des protéines interviennent dans plusieurs fonctions différentes.

Prenons par exemple une protéine ayant les fonctions A et B. D'un côté, on peut estimer qu'il s'agit de fonctions alternatives, c'est-à-dire que l'on étudie parfois la protéine de fonction A, et d'autres fois la protéine de fonction B. Dans ce cas, il faudrait considérer deux objets différents qui seraient classés séparément, et apparaîtraient donc à deux endroits différents de la hiérarchie.

Mais d'un autre côté, on peut vouloir exprimer le fait que cette protéine a une capacité particulière avec deux fonctions différentes, par exemple dans le cas de la présence ou de l'absence d'un partenaire. Dans ce cas, la représentation de l'organisation des protéines doit tenir compte de cette multifonctionalité. Pour représenter ceci, l'utilisation

du modèle pyramidal est très utile. Les pyramides permettent en effet de représenter des classes recouvrantes qui sont rarement étudiées en général [Palla *et al.*, 2005], mais sont pourtant importantes.

Le découpage automatique que nous avons développé permet de réaliser notamment des classifications mixtes hiérarchiques-pyramidales. Nous avons mis en pratique cette méthode de classification avec des données transcriptome, permettant ainsi de faire apparaître des relations entre les gènes et de les classer. Cette méthode de classification peut s'appliquer de manière beaucoup plus générale à d'autres types de données. En effet, la classification est basée sur la définition d'une distance entre des individus. Par conséquent, cette méthode peut être utilisée dans des domaines bien différents, comme par exemple des groupes de consommateurs dans un supermarché, ou bien encore des groupes de joueurs de tennis ayant des profils de victoires plus ou moins similaires. Dans notre cas, il est clair que cette méthode peut s'appliquer à d'autres espèces ou d'autres types de données. De plus, dans le cas de l'utilisation que nous en avons faite sur les données transcriptome, il pourrait être intéressant d'enrichir la distance entre les gènes par une annotation fonctionnelle, en plus des données d'expression. Différents types d'information peuvent être combinés de manière à obtenir une distance plus complète avant de réaliser la classification.

Certes, le découpage automatique n'est pas très robuste en l'état. Il pourrait être intéressant, plutôt que de qualifier les partitions de plus haut niveau, d'étudier plus généralement la fonction qualifiant les partitions de manière globale, et de chercher un extremum en utilisant par exemple des méthodes numériques (Monte-Carlo, recuit simulé).

Par ailleurs, il faut noter que la structure hiérarchique même n'est pas très robuste. Ainsi, de faibles variations des données transcriptome conduisent à des structures différentes. Or, ces données sont largement bruitées. Par conséquent, il pourrait être intéressant de considérer une approche de construction de multiples structures hiérarchiques. Par similarité avec les études phylogénétiques, une structure consensus pourrait ensuite être construite de manière à refléter une structure plus générale et plus robuste des données [Clauset *et al.*, 2008]. De plus, les pyramides pourraient être remplacées par des méthodes permettant des recouvrements multiples, comme par exemple les treillis de Galois.

Concernant la méthode d'identification automatique de biais que nous avons développée et implémentée sous la forme de l'outil BiasSeeker, plusieurs aspects pourraient être approfondis. La limite principale de cette méthode est que nous ne tenons compte ni des niveaux d'induction des gènes, ni de la relation quantitative entre les ARNm et les protéines. Une information sur les concentrations relatives serait intéressante, mais elle n'est pas directement disponible par les données transcriptome. D'autre part, les dépendances entre les différents biais pourraient être explicitées, afin de développer un modèle plus réaliste, et de prendre en compte les tests statistiques multiples et leurs dépendances. L'exploitation de ce travail est en cours et elle devrait permettre de rendre cet outil disponible. Il pourrait aussi être intéressant de généraliser l'approche en consi-

dérant non plus deux échantillons mais plusieurs. Pour cela, le test de Wilcoxon pourrait être remplacé par le test de Kruskal-Wallis.

Les méthodes de prédiction d'interactions protéine-protéine que nous avons développées nous ont permis de construire un réseau d'interactions chez *Synechocystis*. Pour toutes les interactions prédites, nous avons étudié si certains critères étaient vérifiés, comme la similarité fonctionnelle des protéines en interaction, ou la présence de domaines connus pour interagir. Cette analyse est intéressante mais reste qualitative, chaque interaction prédite vérifiant ou non les critères étudiés. Il serait intéressant de réaliser une évaluation quantitative des interactions prédites, de manière à les classer toutes les unes par rapport aux autres. Goldberg *et al.* ont proposé d'évaluer les interactions protéine-protéine en se basant sur l'hypothèse que le réseau vérifie certaines propriétés topologiques (small-world) [Goldberg et Roth, 2003]. Ils utilisent ainsi un coefficient, appelé coefficient hypergéométrique, qui permet d'évaluer dans quelle mesure chaque interaction correspond au modèle défini. Néanmoins, les caractéristiques topologiques globales des réseaux restent des hypothèses fortes qui ne sont pas démontrées et sont remises en question [Han *et al.*, 2005], [Yu *et al.*, 2008].

La méthode *InteroBH* possède l'avantage d'être assez flexible, puisque différents seuils de confiance peuvent être définis. Ceci pose la question du compromis entre la spécificité et la sensibilité. En effet, si on choisit un seuil très strict, on limite le nombre de faux positifs et on augmente par conséquent la spécificité de la méthode de prédiction. Par contre, on limite alors le nombre d'interactions prédites, ce qui diminue sa sensibilité.

La méthode *InteroPorc* est quant à elle moins flexible, mais permet une automatisation. Toutefois, il faut noter que le transfert ne peut être effectué que pour les protéines très conservées entre les espèces. La sensibilité de cette méthode est donc plus limitée. Cela dit, dans le cadre d'une approche à grande échelle et automatique telle que nous l'avons adoptée pour l'implémentation de l'outil *InteroPorc*, il est important de limiter les faux positifs.

Il conviendrait ici de faire des estimations des taux de faux positifs et faux négatifs de manière à étudier les performances de ces méthode de manière quantitative.

Outre la méthode utilisée, la qualité des interactions prédites dépend de la qualité des interactions sources transférées. Si les jeux de données récents semblent de très bonne qualité [Yu *et al.*, 2008], ce n'est pas nécessairement le cas de tous les premiers jeux de données pour lesquels de forts taux de faux positifs ont été mis en évidence [Huang *et al.*, 2007a]. Ainsi, il est essentiel de choisir les données sources en fonction du problème considéré. C'est pourquoi la possibilité d'utiliser l'outil *InteroPorc* avec un jeu de données quelconque pour les interactions sources est très importante. Pour qualifier les interactions prédites, il faudrait alors prendre en compte non seulement les paramètres du transfert, comme cela a été fait pour la méthode *InteroBH* avec la E-value jointe, mais aussi la qualité de l'interaction source. Il manque à ce niveau des indicateurs qualitatifs ou quantitatifs pour caractériser les interactions mises en évidence expérimentalement.

Même si les seuils sur la similarité de séquence sont très stricts, les relations d'orthologie restent des hypothèses qui ne pourraient être confirmées que par la construction d'arbres phylogénétiques. Les solutions choisies ici restent assez simplistes et ne prennent pas en compte les éventuelles difficultés, comme lorsque la protéine orthologue ayant gardé la même fonction n'est pas celle qui a la plus forte similarité de séquence par exemple. De plus, les transferts sont limités quand il s'agit d'espèces très éloignées, comme des eucaryotes supérieurs et des bactéries par exemple.

Par ailleurs, nous avons utilisé des méthodes de prédiction très génériques, c'est-à-dire qui peuvent s'appliquer à toutes les espèces. Ceci nous a permis par la suite d'implémenter un outil automatique applicable à toutes les espèces séquencées. Toutefois, il serait intéressant de profiter de certaines spécificités de *Synechocystis* afin de compléter ce premier réseau. Par exemple, certaines méthodes tiennent compte de la synthénie chez les procaryotes (fusion de gènes, structure d'opérons).

Concernant l'outil de prédiction d'interactions protéine-protéine *InteroPorc* que nous avons développé, il faut noter qu'il est limité de par sa nature aux interactions déjà connues dans d'autres espèces. C'est en effet le principe de la méthode de prédiction de transférer des interactions d'une espèce vers une autre en utilisant les interologues. Par conséquent, cette méthode permet de prédire des interactions nouvelles pour une espèce, mais pas d'interactions spécifiques à celle-ci. L'espace exploré reste celui des interactions déjà mises en évidence expérimentalement chez au moins une espèce. En revanche, cette méthode de prédiction permet de transférer de manière rapide la connaissance actuelle sur différentes espèces, à une espèce non encore explorée, et notamment les espèces nouvellement séquencées.

Deux autres approches ont été adoptées dans d'autres travaux pour construire des réseaux plus larges de relations fonctionnelles. D'un côté, la base de données STRING met à disposition des prédictions provenant de la co-expression, la co-localisation, la co-citation dans la littérature et l'identification expérimentale [Jensen *et al.*, 2008]. L'interface web associée permet d'explorer les interactions prédites pour certaines protéines. Néanmoins, il est beaucoup plus difficile d'obtenir des résultats complets à l'échelle d'une espèce, et les données sources ne peuvent pas être contrôlées. Ainsi, les objectifs sont différents de ceux de l'outil *InteroPorc* qui permet de prédire des interactions physiques à l'échelle d'un protéome à partir de données éventuellement choisies par l'utilisateur.

D'un autre côté, Kim *et al.* ont élaboré une base de données de relations fonctionnelles chez *Synechocystis* [Kim *et al.*, 2008] à partir de quatre bases de données : PSIMAP [Park *et al.*, 2001], iPFAM [Finn *et al.*, 2005], InterDom [Ng *et al.*, 2003b] et STRING [Jensen *et al.*, 2008]. Même si ces relations ne sont pas nécessairement des interactions physiques, il serait intéressant de comparer nos prédictions avec celles-ci.

Jusqu'à présent, l'outil *InteroPorc* fonctionne pour une espèce donnée, de manière globale. Ainsi, il suffit de choisir une espèce et les interactions sources sont transférées sur cette espèce. Néanmoins, on est souvent amené à se poser des questions sur une protéine en particulier, ou un groupe de protéines. Par conséquent, il serait très utile de mettre en place un système d'interrogation centré sur une protéine. Ceci permettrait

d'obtenir la liste des interactions prédites avec cette protéines. De plus, un accès par un service web pourrait également s'avérer utile.

L'analyse des réseaux d'interactions protéine-protéine obtenus par des méthodes d'identification à grande échelle pourrait être enrichie en considérant d'autres espèces, en particulier *Caenorhabditis elegans* et *Drosophilia melanogaster*. D'autre part, il conviendrait de faire une analyse fonctionnelle des différents réseaux d'interactions, afin de caractériser les différences et ressemblances en termes de fonctions biologiques. En effet, ces différentes études restent biaisées même lorsqu'elles sont à grande échelle. Les appâts sont souvent choisis par rapport à l'espèce considérée ou par rapport à certaines problématiques. Ainsi, en plus des limitations techniques des différentes méthodes d'identification, les interactomes ne sont pas explorés de manière uniforme. C'est pourquoi il serait intéressant de quantifier ces différences, notamment entre les différentes espèces.

Finalement, nous avons identifié les modules fonctionnels en nous basant sur les réseaux d'interactions protéine-protéine, mais d'autres structures pourraient être étudiées dans ces graphes, par exemple des voies de signalisation. De plus, la dynamique a été introduite ici au niveau des modules. On pourrait également l'introduire au niveau des protéines et considérer des graphes dont les nœuds sont caractérisés par un état, par exemple actif ou inactif.

Pour étudier les réponses aux stress oxydants et aux métaux lourds, nous avons fait une analyse de la réponse transcriptionnelle au niveau des gènes. Nous avons identifié et classé les gènes qui répondent par un signal semblable, c'est-à-dire ceux qui sont co-exprimés. Le but est de regrouper les gènes qui fonctionnent ensembles pour la réalisation d'un processus biologique. On imagine alors que ceci provient d'un stimulus particulier. Néanmoins, certains gènes peuvent envoyer un signal similaire, sans pour autant répondre au même stimulus. Il est alors intéressant de compléter l'information sur la régulation de la transcription par une information sur les contacts physiques entre les protéines. Ainsi, cette information complémentaire permet d'enrichir la vision du phénomène global et potentiellement de distinguer les cas où les gènes répondent à un même stimulus ou pas. En d'autres termes, ceci peut permettre de distinguer les liens de corrélation des liens de causalité. En effet, les protéines codées par les gènes co-exprimés et qui sont en interaction ont de fortes chances de réaliser un processus biologique. Par conséquent, les réseaux d'interactions protéine-protéine permettent de préciser la prédiction des réseaux de régulation réalisée à partir de données d'expression. Par ailleurs, l'intégration des ces deux types de données pourrait permettre une prédiction conjointe des réseaux de régulation et des réseaux d'interactions protéine-protéine.

Conclusion

Une vérité importante naissait dans ma tête : une vie scientifique pouvait être intéressante socialement et intellectuellement.

James D Watson,
The double helix, 1968

À l'interface de la biologie et de la bioinformatique, ce travail a donné lieu au développement de méthodes et d'outils pour étudier entre autres les réponses cellulaires aux stress oxydants et aux métaux lourds. Nous avons mis en évidence en particulier quatre résultats principaux.

Tout d'abord, un découpage automatique de hiérarchie a été développé. Il nous a permis de réaliser une classification mixte hiérarchique-pyramidale dont l'un des avantages est d'être très générale. En effet, cette classification est seulement basée sur une distance entre des objets quelconques.

De plus, une méthode d'identification de biais de composition entre deux groupes de protéines a été réalisée. Cette méthode présente l'avantage d'être automatique et très générale puisqu'elle peut s'appliquer à deux groupes quelconques de protéines. Dans notre cas, nous avons considéré des groupes provenant de données d'expression. En outre, nous avons développé un outil, BiasSeeker, permettant de mettre en œuvre cette méthode de manière simple, rapide et visuelle.

Par ailleurs, un outil automatique de prédiction d'interactions protéine-protéine a été développé. Il nous a permis entre autres de construire un réseau d'interactions protéine-protéine chez la cyanobactérie *Synechocystis*. Cet outil, *InteroPorc*, possède les quatre avantages suivants : il est rapide, puisque des prédictions à l'échelle d'une espèce sont réalisées en trois minutes environ ; il est automatique et général, car il peut s'appliquer à toutes les espèces dont le génome est séquencé ; il est flexible, dans la mesure où la méthode de prédiction a été implémentée indépendamment des données utilisées, ce qui permet notamment de sélectionner les interactions sources qui vont être transférées ; il est disponible pour tous, grâce à une interface web où il peut être utilisé en ligne ou téléchargé pour une utilisation en local.

Enfin, différents formalismes ont été proposés pour l'étude des interactions protéine-protéine et de leur dynamique. Certains formalismes permettent par exemple de prendre en compte des indicateurs techniques de manière automatique pour l'études des interactions protéine-protéine. D'autres permettent d'étudier les groupes de protéines présents

dans les réseaux et notamment leur dynamique en se basant sur des données d'expression. Même si ces formalismes sont simples, ils ouvrent la porte à des analyses automatiques.

Ainsi, ces méthodes nous ont permis d'étudier la régulation de la transcription en termes de relations fonctionnelles entre les gènes, puis d'y ajouter les contacts physiques entre les protéines correspondantes grâce à des méthodes de prédiction d'interactions protéine-protéine.

Dans un contexte multi-disciplinaire comme celui de la bioinformatique, l'un des principaux défis est de combiner plusieurs approches, plusieurs expertises, plusieurs cultures, afin d'aller plus loin dans l'analyse et la compréhension du vivant. Ceci nécessite de comprendre les stratégies, les avantages et les limites des autres approches. Des difficultés peuvent être rencontrées notamment quand les interactions entre les différents experts ne sont pas suffisamment fréquentes ou efficaces. Ceci peut entre autres conduire à concevoir des expériences avant de définir le formalisme qui sera utilisé pour les analyser. Inversement, en amont du projet, il arrive qu'un formalisme soit choisi ou défini sans que les éléments clefs de la problématique biologique ne soient pris en compte de manière pertinente. Il est donc primordial d'unifier les différentes cultures le plus tôt possible pour avancer ensemble vers un objectif commun.

Cette question de l'interdisciplinarité est essentielle et non triviale. Ainsi, Thomas Cech et Gerald Rubin ont proposé la mise en place d'un lieu de recherche conçu pour favoriser les contacts entre les chercheurs et faire tomber les murs qui existent entre les disciplines [Cech et Rubin, 2004]. Si cette collaboration est évidemment difficile, c'est aussi un des intérêts majeurs de la discipline. C'est ce défi qui rend la bioinformatique si passionnante.

Troisième partie

Annexes

Annexe A

Profils d'expression de familles de gènes

Strain and treatment	WT : Cd WT : SC										1738:Cd WT : Cd	WT : H2O2 WT : SC					WT : Fe- WT : SC		WT : Fe+ WT : SC		WT : Zn+ WT : SC		Gene name
Time or C	15'	30'	60'	75'	90'	180'	300'	300b'	360'	960'	180'	15'	30'	180'	300'	420'	1-0	2-0	240'	360'	30'	240'	
A: PROTEIN SYNTHESIS, FOLDING, AND TURNOVER																							
Chaperones																							
sll0430	0,56	1,81	**2,04**	**2,18**	**6,55**	**4,75**	**2,94**	**4,33**	1,73	1,10		1,37	**0,44**	1,56	1,22	1,47	0,58	**0,38**	**2,67**	**2,41**	1,05	1,82	htpG
sll1514	0,55	**3,07**	**4,24**	**3,42**	**13,41**	**14,53**	**13,47**	**13,48**	**10,34**	**3,50**	**0,42**	**4,36**	**9,16**	**17,45**	**9,52**	**16,14**	0,65	**0,07**	**2,56**	**3,08**	1,78	**3,95**	hspA
sll0170	0,60	**2,07**	**2,10**	**2,01**	**4,97**	**3,26**	**2,01**	**2,32**	**2,64**	1,75		**2,11**	1,19	**4,26**	**1,78**	**2,35**	0,64	0,54	1,58	1,67	1,09	**2,96**	dnaK
sll0416	0,68	**2,15**	**2,01**	**1,97**	**3,83**	**3,79**	**3,19**	**3,81**	**2,43**	1,00		0,77	**0,46**	1,00	1,76	1,33	**0,46**	0,67	1,40	1,67	0,69	**2,61**	groEL2
sll0897 [a]	1,07	1,33	1,14	1,19	**1,97**	**2,54**	**2,29**	**2,57**	**2,50**	1,08		**0,46**	1,58	**2,28**	1,41	1,24	0,86	0,97	1,01	1,34	1,11	**2,37**	dnaJ
ssll1707 [a]	0,89	1,26	1,21	0,98	1,81	**2,43**	**2,16**	**2,66**	**2,34**	1,04		**0,44**	1,53	**2,03**	1,33	1,46	0,75	1,11	1,12	1,51	0,98	**2,44**	hs
sll1933					1,01		1,22	1,85	1,29	1,17		0,55	**0,33**	1,15	0,88	1,29	0,81	1,16	1,21	1,13	0,63	1,44	dnaJ
sll1666	1,12	1,18	1,26	1,54	**2,10**	**2,10**	1,75	1,81	1,38	1,46		0,73	0,79	0,91	0,74	0,81	1,06	1,20		0,92		1,82	dnaJ-like
slr0093	1,05	**2,02**	1,53		**2,71**	**3,92**	**2,53**	**3,84**	**2,95**	1,54		1,84	**9,38**	**9,15**	**2,95**	**6,61**	0,76	**0,51**	1,12	1,14	1,01	1,18	dnaJ
slr2075	0,74		**2,14**	1,77	**3,40**	**2,78**	**2,25**	**2,63**	**2,06**	1,08		1,06	**0,42**	1,08	1,29	1,86	**0,33**	**0,41**	1,34	1,39		1,84	groES
slr2076	0,60	1,72	**2,25**	1,80	**3,70**	**2,99**	**2,33**	**2,51**	**2,24**	1,42		1,15	**0,31**	1,19	1,03	1,68	**0,40**	**0,51**	1,57	0,95	1,02	**2,13**	groEL1
sll0057	1,03	0,95	1,04	1,02	1,13	1,23	1,15	1,37	1,50	0,99		**0,41**	**0,44**	1,09	1,10	0,86	0,64	0,82	1,36	1,61	0,63	1,20	grpE
Degradation of proteins, peptides, and glycopeptides																							
slr1204	1,13	**26,62**	**18,87**	**12,58**	**22,63**	**4,80**	**7,07**	**9,27**	**7,71**	**2,29**		**2,93**	**2,03**	1,37	1,73	**7,07**	0,67	1,05	**0,51**	0,75	**2,86**	0,95	htrA
slr1641	**0,45**	**1,96**	1,10	0,98	**4,75**	**6,88**	**3,78**	**6,08**	**4,30**	1,20	**0,49**	**2,88**	0,61	**11,40**	**3,33**	**4,58**	0,79	0,59	1,26	1,28	0,76	**8,18**	clpB2
slr1751	1,68	**2,14**	1,89	1,76	**4,24**	**3,71**	**3,58**	**4,61**	**4,66**	**1,99**		1,48	1,71	1,83	0,61	1,06	0,76	0,68	0,99	0,70	0,79		clpC
sll0020	0,75	1,43	1,35	1,57	**3,03**	**2,37**	**2,62**	NS	**2,34**	1,13	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	clpB
slr0257		1,33	1,41	1,51	**2,82**	**2,13**	**3,08**	**2,44**	**2,28**	1,32		0,61	1,22	1,28	0,97	0,91	0,79	0,76	1,08	1,10	0,80	1,36	clpB
sll1427	1,35	1,66	1,56	1,80	**3,12**	**2,67**	**2,33**	**3,13**	**2,26**	1,61		1,06	0,64	1,23	0,77	1,11	0,92	1,09	0,86	0,73	0,85	**1,96**	hhoB
sll1679	1,30	1,81	1,76	1,72	**3,31**	**3,65**	**2,57**	**3,74**	1,86	1,85		**0,42**	**0,31**	0,88	0,60	0,80	0,64	1,11	0,97	0,72	0,97	1,39	hhoA
slr0165	1,08	1,12	1,20	1,56	1,85	1,34	1,05	1,03	0,78	0,97		**0,49**	**0,49**	0,89	0,96	0,82	0,61	0,83	1,60	1,38	0,98	1,05	clpP3
slr0542	1,19	1,00	1,27	1,15	1,39	1,41	1,37	1,51	1,34	0,95		**0,51**	**0,49**	1,14	1,40	0,78	0,70	0,70	1,68	1,52	0,74	1,15	clpP1
sll0535	1,03	0,94	0,89	0,93	0,96	0,96	0,98	1,59	0,90	0,90		**0,36**	1,25	1,21	0,83	1,13	0,92	0,89	1,19	1,01	1,22	1,77	clpX
sll1343	0,86	1,11	1,03	1,10	0,94	0,90	0,90	0,74	0,95	1,01		**0,32**	0,91	0,98	0,90	0,91	1,01	1,22	0,97	0,94	0,74	1,60	pepN
slr0008	1,32	1,31	1,09	1,02	1,15	1,41	1,08	1,39	1,74	0,98		1,08	**2,71**	**2,64**	1,01	1,45	**0,49**	0,66	1,04	0,79	1,19	**1,97**	ctpA
sll1703		1,03	0,89	0,85	1,03	1,07	0,97	1,64	0,83	0,98		**0,53**	0,55		0,89	1,14	0,94	0,96	0,99	0,99	1,14	**2,52**	sppA
Protein modification and translation factors																							
slr0885 [b]		1,21	1,41	1,53	**2,99**	**2,87**	**3,01**		**2,42**	1,37		**10,23**		**2,50**	0,79	1,58	0,96			0,96		0,85	hs
sll0869 [b]		1,17			**1,98**	**1,94**	**2,68**	**2,75**	**1,91**	1,25		**2,69**	1,78					1,28	**0,21**	0,89		**1,98**	est
sll0868 [b]	0,99	1,19	1,21	1,34	1,79	1,53	1,32	**1,97**		1,06		0,91	**0,52**	1,46	0,88	1,27	1,02	1,01	1,08	0,88	1,11		lipA
sll0867 [b]	1,06	1,20	1,44	1,41	**2,15**	1,78	1,83	1,71	1,15	1,25		0,72	0,82	1,19	0,81	0,93	0,93	1,12	**2,38**	1,11	1,57	0,60	hs
slr0974	0,99	1,09	1,23	0,71	1,64	**2,13**	1,62	**2,02**	**2,92**	1,09		**0,35**	1,17	1,59	1,09	1,41	0,99	1,05	1,11	1,45		0,83	infC
sll0145	0,91	0,78	0,90	0,79	**0,48**		**0,47**	**0,46**	**0,44**	0,72		**0,52**	**0,48**	0,64	0,88	0,70	1,09	1,25	1,39	1,33	0,54	**0,51**	rrf
slr0434	1,12	0,86	0,88	0,78	**0,51**	**0,51**	**0,46**	**0,48**	**0,42**	0,79	**2,10**	**0,52**	**0,35**	0,53	0,80	1,00	0,98	1,38	1,39	1,33	1,05	0,55	efp
sll0546	0,94	0,94	0,85	0,87	0,72	**0,48**	0,54	0,63	0,67	0,88		0,88	0,73	0,70	1,04	1,34	1,23	0,79	1,01	1,11	**0,53**	1,01	ycH
slr1251	0,82	0,92	0,82	0,63	**0,37**	**0,45**	0,55	0,55	0,81	0,89		1,10	**0,26**	0,67	1,15		0,76	1,48	1,14	1,45	1,19	0,63	ppiB
sll0830	0,87	0,94		1,02	0,99	0,81	0,90					**7,91**	**2,33**		0,91	0,98		**0,51**	0,89	1,34	0,85		fus
sll1098 [c]	1,06	0,86	0,93	1,15	0,86	**0,55**	**0,52**	1,20	**0,38**	0,79		0,84	**0,68**	1,18	0,83	1,41	0,85	1,04	1,81	0,89	1,04	2,13	fus
sll1099 [c]	1,00	0,92	1,07	0,99	0,86	**0,49**	**0,45**	0,64	**0,42**	1,08		0,61	0,56	0,92	0,93	1,09	0,76	1,11	1,80	1,05	0,95	1,79	tufA
slr1105	1,30	1,44	1,07	0,95	0,86	1,15	1,42	1,25	0,97	0,91		1,18	**3,15**	1,43	1,03	1,18	1,16	1,12	1,55	1,24	1,04	**2,24**	bipA like
sll2001	1,20	1,04	1,23	1,10	1,01	1,19	1,48	1,39	1,25	1,17		**0,49**	**0,34**	1,03	1,09	0,99	0,80	1,69	1,26	0,98		1,10	pepA
sll0555				0,75	0,88	1,17	1,72	1,69				**0,43**	**0,46**	0,65	0,95	0,86	0,59	0,90	1,19	1,08	0,99	1,21	map

FIG. A.1 – Profils d'expression des gènes participant à la synthèse, la maturation et la dégradation des protéines (partie 1/2). Les gènes sont regroupés par famille. Les valeurs de log-ratio entourées indiquent que le gène est induit alors que les valeurs en gras sur des cases grises indiquent que le gène est réprimé. Ceci permet d'avoir une vision globale de la réponse des gènes pour les différentes conditions.

Strain and treatment	WT : Cd / WT : SC										1738:Cd / WT : Cd	WT : H2O2 / WT : SC					WT : Fe- / WT : SC		WT : Fe+ / WT : SC		WT : Zn+ / WT : SC		Gene name
Time or C	15'	30'	60'	75'	90'	180'	300'	300b'	360'	960'	180'	15'	30'	180'	300'	420'	1-0	2-0	240'	360'	30'	240'	
A: PROTEIN SYNTHESIS, FOLDING, AND TURNOVER																							
Ribosomal proteins: synthesis and modification																							
sll1096	NS	NS	NS	NS	NS	NS	NS	0,66	NS	NS		0,75	**0,36**	0,83	1,34	0,99	0,79	1,14	1,70	**2,07**	1,04	1,15	rps12
sll1097	0,85	0,95	0,89	0,77	0,65	0,57	0,64	0,59	0,76	0,94		1,05	**0,52**	0,99	1,47	1,30	0,84	0,99	1,84	1,67	1,20	0,98	rps7
sll1101	1,09	0,92	1,05	0,90	0,71	**0,38**	**0,33**	**0,44**	**0,38**	0,96	**2,11**	**0,33**	**0,19**	0,86	1,65	1,07	0,70	1,06	1,53	1,53	0,90	0,73	rps10
sll1244	0,89	1,09	0,92	0,71	0,55	**0,48**	**0,49**	**0,52**	**0,35**	0,84	**2,07**	0,61	**0,40**	0,75	1,06	0,93	0,58	0,66	1,50	1,19	0,80	**0,36**	rpl9
sll1260	0,85	0,91	0,99	0,86	0,62	**0,47**	**0,48**	0,54	0,60	1,00	**2,22**	0,92	**0,25**	1,10	1,42	1,44	**0,46**	1,23	1,31	1,61	0,80	**0,46**	rps2
sll1740	0,82	1,05	0,99	0,94	0,92	0,74	0,99	0,76	1,17	1,05		0,81	**0,40**	0,99	1,18	1,17	0,78	0,85	1,10	1,05	1,04	0,89	rpl19
sll1744 [d]	0,97	0,85	0,98	0,80	0,67	**0,49**	0,56	**0,46**	0,74	0,79		1,37	0,70	1,38	1,56	1,76	0,87	0,82	1,12	1,41	0,65	0,60	rpl1
sll1745 [d]	1,04	0,78	1,06	0,97	1,40	0,53	**0,45**	**0,51**	0,62	0,76	**3,23**	0,65	0,74	0,69	1,62	0,99	0,72	1,37	1,08	1,41	0,73	**0,52**	rpl10
sll1746 [d]	0,99	0,82	0,93	0,85	0,71	0,54	**0,50**	**0,36**	0,57	0,86	**2,71**	1,26	1,10	0,65	1,67	0,99	0,57	1,33	0,83	1,52	1,09		rpl12
sll1799 [e]	1,07	0,88	0,67	0,63	**0,34**	**0,24**	**0,28**	**0,24**	**0,32**	0,83	**3,66**	0,76	0,74	0,81	1,46	1,67	1,13	0,86	1,05	1,39	0,66	0,65	rpl3
sll1800 [e]	NS	NS	NS	NS	NS	NS	NS	**0,25**	NS	NS	**3,73**	**0,32**	0,72	0,65	1,66	1,27	0,97	0,89	1,08	1,74	**0,48**	0,75	rpl4
sll1801 [e]	1,27	0,79	0,90	0,66	**0,31**	**0,21**	**0,22**	**0,22**	**0,26**	0,88	**3,39**	0,73	0,95	0,77	1,51	1,43	0,98	1,07	1,09	1,41	0,65	0,81	rpl23
sll1802 [e]	1,44	0,86	0,93	0,65	**0,45**	**0,36**	**0,34**	**0,27**	**0,42**	0,84	**3,04**	**0,45**	0,68	0,73	1,36	1,35	0,83	1,06	1,02	1,53	0,71	**0,45**	prl2
ssl3432 [e]	NS	NS	NS	NS	NS	NS	NS	**0,24**	NS	NS	**2,86**	**0,40**	1,11	0,58	**2,37**	0,94	0,64	0,76	1,06	**2,07**	0,55	**0,48**	rps19
sll1803 [e]	1,19	0,93	0,86	0,77	**0,47**	**0,26**	**0,27**	**0,28**	**0,32**	0,79	**3,49**	**0,46**	0,88	0,73	1,47	1,30	0,75	0,80	1,18	1,55	0,70	0,76	rpl22
sll1804 [e]	1,11	0,84	0,93	0,70	**0,40**	**0,25**	**0,26**	**0,25**	**0,28**	0,79	**2,94**	**0,34**	0,56	0,61	1,54	1,02	0,88	1,33	1,23	1,74	0,57	0,75	rps3
sll1805 [e]	1,26	0,84	0,97	0,68	**0,38**	**0,30**	**0,26**	**0,29**	**0,30**	0,84	**3,15**	**0,46**	**0,48**	0,75	1,50	1,21	0,78	0,87	1,16	1,54	0,76	0,57	rpl16
ssl3436 [e]	1,08	0,93	1,05	0,73	**0,49**	**0,23**	**0,20**	**0,21**	**0,22**	0,79	**3,20**	**0,51**	0,69	0,65	1,41	1,12	0,79	0,95	1,12	1,56	0,73	**0,28**	rpl29
ssl3437 [e]	1,13	0,89	1,03	0,68	**0,36**	**0,21**	**0,19**	**0,23**	**0,21**	0,81	**3,13**	**0,38**	0,64	0,61	1,40	0,93		1,11	1,09	1,50	0,77	**0,46**	rps17
sll1806 [e]	NS	NS	NS	NS	NS	NS	NS	**0,19**	NS	NS	**2,76**	0,59	0,64	0,57	1,52	1,00	0,70	1,01	1,63	1,71		**0,34**	rpl14
sll1807 [e]	1,27	0,83	1,07	0,64	**0,36**	**0,23**	**0,18**	NS	**0,23**	0,78	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	rpl24
sll1808 [e]	NS	NS	NS	NS	NS	NS	NS	**0,23**	NS	NS	**2,76**	**0,41**	**0,49**	0,55	1,41	1,00	0,67	**2,23**	1,26	1,67	0,74	**0,37**	rpl5
sll1809 [e]	0,97	0,95	1,06	0,72	**0,48**	**0,37**	**0,36**	**0,18**	**0,41**	0,81	**3,08**	**0,42**	0,63	0,55	1,65	0,97	0,57	0,89	0,92	1,71	0,74	**0,47**	rps8
sll1810 [e]	0,93	0,96	1,04	0,79	**0,48**	**0,26**	**0,24**	**0,41**	**0,30**	0,71	**2,18**	0,68	**0,48**	0,80	1,18	1,37	0,78	0,87	1,07	1,45	0,84	1,32	rpl6
sll1811 [e]	NS	NS	NS	NS	NS	NS	NS	**0,28**	NS	NS	**2,33**	1,03	0,62	0,65	1,41	1,11	0,86	0,98	1,03	1,34	1,13	**0,31**	rpl18
sll1812 [e]	NS	NS	NS	NS	NS	NS	NS	**0,36**	NS	NS	**2,31**	0,70	**0,47**	0,80	1,22	1,13	0,73	0,88	1,08	1,41	1,24	**0,50**	rps5
sll1813 [e]	NS	NS	NS	NS	NS	NS	NS	**0,24**	NS	NS		0,91	**0,37**	0,65	1,08	1,26	0,87	0,90	1,12	1,28	1,28	0,88	rpl15
sll1816 [f]	NS	NS	NS	NS	NS	NS	NS	**0,52**	NS	NS		1,07	**0,28**	0,99	1,38	1,46	0,91	0,92	1,49	1,54	1,05	0,82	rps13
sll1817 [f]	NS	NS	NS	NS	NS	NS	NS	**0,43**	NS	NS		0,70	**0,30**	0,68	1,58	1,21	0,91	1,00	1,53	**1,96**	0,84	0,87	rps11
sll1819 [f]	1,15		0,92	0,77	0,70	0,64	0,68	**0,52**	0,73	1,00		0,98	0,79	0,74	1,25	0,95	0,77	0,98	1,33	1,36	1,16	**0,52**	rpl17
sll1821 [f]	1,02	0,84	1,06	0,78	0,75	**0,51**	**0,50**	**0,44**	**0,38**	0,85		**0,49**	**0,32**	0,77	1,05	0,96	0,97	1,14	1,35	1,37	0,89	1,18	rpl13
sll1822 [f]	1,15	0,90	1,01	0,85	0,71	0,53	**0,54**	0,69	**0,45**	0,86		0,58	**0,31**	0,90	0,97	1,03	0,71	0,81	1,49	1,46	0,87	1,15	rps9
ssl3445 [f]	1,17	0,89	0,97	0,66	**0,48**	**0,42**	**0,44**	**0,41**	**0,33**	0,94		**0,52**	0,58	0,70	1,40	0,83	0,57	0,85	1,23	1,25	0,72	0,83	rpl31
sll1824 [f]	0,99	0,86	0,89	0,85	0,60	**0,49**	**0,48**	**0,49**	**0,32**	0,92		0,84	**0,27**	0,88	0,96	1,15	0,76	0,91	**1,99**	1,27	1,05	1,14	rpl25
smr0011 [g]		1,05	0,89	0,68	**0,42**	0,56	0,65	0,59	0,68	0,89		1,57	1,38	0,87	1,70	0,99	0,76	0,67	1,18	1,40	0,85	0,71	rpl34
slr1469 [g]	0,99	0,85	0,91	0,76	**0,47**	**0,46**	**0,52**	0,57	0,59	0,98		1,20	0,77	0,92	1,04	1,31	0,68	0,70	1,47	1,21	0,92	0,74	rnpA
slr1470 [g]	0,92	0,93	1,00	1,06	0,88	0,64	0,54	0,60	**0,47**	0,80		1,39	0,98	1,19	0,93	1,28	0,86	0,66	1,56	1,15	0,88	0,77	ho
sll0767 [h]	0,83	0,84	1,03	0,99	0,68	**0,45**	0,60	0,54	0,56	0,91		0,66	**0,42**	0,84	1,33	0,95	0,95	1,05	1,36	1,08	0,71	0,85	rpl20
ssl1426 [h]	0,83	0,84	0,96	0,81	**0,47**	**0,40**	**0,48**	**0,53**	**0,45**	1,24	**2,60**	1,05	**0,39**	0,91	1,19	1,15	0,79	0,93	1,32	1,16	0,95	0,57	rpl35
ssr0482	0,75	0,89	0,84	0,67	**0,43**	0,56	0,80	0,60	1,14	0,93		0,86	**0,41**	0,77	1,16	0,88	1,79	1,90	0,71	0,90	0,94	0,96	rps16
ssr1398	NS	NS	NS	NS	NS	NS	NS	**0,35**	NS	NS	**2,55**	0,68	**0,35**	0,70	1,18	0,71	0,68	0,87	**4,32**	1,29	0,84	**0,23**	rpl33
ssr1399	NS	NS	NS	NS	NS	NS	NS	**0,30**	NS	NS		1,21	**0,50**	0,77	1,55	0,80	0,72	0,63	1,10	1,32	0,78	**0,39**	rps18
ssr1604	0,93	0,95	0,92	0,86	**0,48**	**0,52**	0,66	0,74	0,68	0,91		0,94	0,54	1,05	1,04	1,47	1,13	0,84	0,89	1,07	0,68	0,64	rpl28
ssr1736	0,94	0,91	0,88	0,72	**0,36**	**0,45**	**0,48**	**0,50**	0,53	0,83	**2,08**	0,86	**0,43**	0,65	1,21	1,14	0,72	1,14	1,90	1,26	0,57	0,75	rpl32
ssr2799	NS	NS	NS	NS	NS	NS	NS	**0,49**	NS	NS		1,81	1,46	0,95	1,03	**0,46**	1,17	1,39	0,88	0,69	1,04	0,59	rpl27
sll1909		0,95	0,77		**0,51**	0,67	0,77	0,75	0,83	0,98		**0,42**	**0,44**	0,64	1,13	0,76	1,00	1,02	1,32	1,67	0,76	0,97	prmA like
slr1356	1,14	1,00	0,83	0,79	0,66	0,62	0,79	0,89	0,99	0,90		**0,47**	**0,47**	0,69	1,19	1,03	1,29	1,24	1,26	1,12	0,90	1,59	rps1a
sll0754	0,76	1,44	1,26	1,79	**3,55**	**3,22**	**2,97**	**2,07**	**3,23**			0,80	0,87	1,13	1,43	0,68	0,62	0,53	1,27	1,65	0,66	0,99	rbfA
slr0852 [i]	0,85	**2,43**	1,38	1,29	**4,20**	**4,22**	**3,76**	**3,87**	**4,32**	1,57		0,97	**12,54**	**3,53**	**1,71**	**2,08**	0,93	**0,50**	1,67	**2,05**	0,95	**2,45**	ho
slr0853 [i]	0,89	**1,95**		1,15	**2,73**	**3,80**	**3,63**	**5,92**	**3,98**			0,89	**4,44**	**4,81**	**2,10**	**2,67**	0,89	**0,54**	1,14	**1,97**	0,98	**3,48**	rimI

FIG. A.2 – Profils d'expression des gènes participant à la synthèse, la maturation et la dégradation des protéines (partie 2/2). Les gènes sont regroupés par famille. Les valeurs de log-ratio entourées indiquent que le gène est induit alors que les valeurs en gras sur des cases grises indiquent que le gène est réprimé. Ceci permet d'avoir une vision globale de la réponse des gènes pour les différentes conditions.

Annexe B

Propriétés des acides aminés

B.1 Liste des acides aminés

Nom	Abréviations	
	3 lettres	1 lettre
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Acide aspartique	Asp	D
Cystéine	Cys	C
Acide glutamique	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Méthionine	Met	M
Phénylalanine	Phe	F
Proline	Pro	P
Sérine	Ser	S
Thréonine	Thr	T
Tryptophane	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V
Quelconque	-	Z

Tab. B.1 – Liste des acides aminés.

B.2 Calcul des paramètres étudiés dans BiasSeeker

L'objectif de cette annexe est de présenter dans le détail les différents paramètres étudiés lors de la recherche automatique de biais effectuée par l'outil BiasSeeker. Pour cela, nous donnons la liste des paramètres en indiquant leur signification et le calcul permettant d'obtenir leur valeur. Afin de préciser les calculs effectués, nous avons considéré une séquence S constituée de L acides aminés dénommés aa_i.

Nous avons classé les différents paramètres étudiés en quatre classes selon qu'ils caractérisaient les propriétés générales, les propriétés des acides aminés, la composition en acides aminés ou encore la composition en atomes des protéines.

B.2.1 Caractérisation des propriétés générales

Nous avons choisi de caractériser une protéine de façon globale en considérant les six catégories suivantes : la longueur ; le poids moléculaire ; le caractère aliphatique ; le caractère hydrophobe ou hydrophile ; la stabilité ; les propriétés optiques.

- *La longueur* : il s'agit du nombre d'acides aminés qui constituent la protéine.
- *Le poids moléculaire* : il s'agit du poids total de la protéine que nous avons calculé selon l'équation B.1 à partir des poids des acides aminés.

$$P = \sum_{i=1}^{L} P(aa_i) \tag{B.1}$$

  où $P(aa_i)$ est le poids de l'acide aminé i.
- *Le point isoélectrique* : il s'agit de la valeur de pH pour laquelle la protéine est neutre, c'est-à-dire sans charge globalement.
- *L'indice aliphatique* : cet indice reflète le caractère aliphatique des acides aminés de la protéine. En chimie organique, un composé aliphatique est un composé carboné cyclique ou acyclique, saturé ou non, à l'exclusion des composés aromatiques (c'est-à-dire qui contiennent un système cyclique respectant la règle d'aromaticité de Hückel). Ce terme était à l'origine utilisé pour décrire les acides gras dont la structure est en chaînes linéaires. L'index aliphatique d'une protéine a été défini par Kyte *et al.* comme le volume relatif occupé par les chaînes aliphatiques [Kyte et Doolittle, 1982] et calculé selon l'équation B.2 :

$$AI = X(Ala) + a * X(Val) + b * \{X(Ile) + X(Leu)\} \tag{B.2}$$

  où X(Ala), X(Val), X(Ile), and X(Leu) sont les proportions par mole des différents acides aminés (alanine, valine, isoleucine, et leucine). Les coefficients a et b sont les volumes relatifs des chaînes de la valine (a=2,9) d'une part et de la leucine et l'isoleucine (b=3,9) d'autre part, par rapport au volume de la chaîne de l'alanine. Cet index peut être vu comme un facteur positif pour la thermostabilité des protéines globulaires.
- *L'indice d'instabilité* : cet indice reflète la nature stable ou instable d'une protéine dans un tube. Une étude statistique de 12 protéines stables et 32 protéines instables a révélé que certains dipeptides apparaissent à une fréquence significativement différente dans les protéines instables en comparaison aux protéines stables [Guruprasad *et al.*, 1990]. Les auteurs ont alors proposé un poids d'instabilité pour chacun des 400 dipeptides (DIWV). En se basant sur ces poids, l'indice d'instabilité a été calculé selon l'équation B.3 :

$$II = \frac{10}{L} * \sum_{i=1}^{L-1} DIWV(x_i x_{i+1}) \tag{B.3}$$

où L est la longueur de la protéine et $DIWV(x_i x_{i+1}$ est le poids d'instabilité pour le dipeptide qui commence à la position i. Une protéine dont l'indice d'instabilité est inférieur à 40 est prédite comme stable. Une valeur supérieure à 40 indique que la protéine est probablement instable.

- *L'absorbance* : l'absorbance est la quantité mathématique qui reflète le phénomène de l'absorption de la lumière. Ce paramètre a été calculé selon l'équation B.4 :

$$A = \frac{E}{P} \tag{B.4}$$

où E est le coefficient d'extinction (voir ci-dessous) et P est le poids moléculaire de la protéine.

- *Le coefficient d'extinction* : ce coefficient indique la quantité de lumière qu'une protéine absorbe à une longueur d'onde donnée. Ce coefficient est utile lorsque l'on veut suivre une protéine par spectrophotomètre lors de sa purification. Il a été montré que ce coefficient d'extinction peut être estimé à l'aide de sa composition en acides aminés [Pace *et al.*, 1995]. Ce coefficient d'extinction a été calculé selon l'équation B.5 :

$$E = Nb(Tyr) * E(Tyr) + Nb(Trp) * E(Trp) + Nb(Cys) * E(Cys) \tag{B.5}$$

où E(Tyr) = 1 490, E(Trp) = 5 500, E(Cys) = 125 (Tyr=tyrosine, Trp=tryptophane, Cys=cystine). Comme la cystéine n'absorbe presque pas pour des longeurs d'onde supérieures à 260 nm alors que la cystine absorbe, deux paramètres différents ont été calculés en se baseant sur deux hypothèses différentes. Selon la première hypothèse, les cystéines apparaissent toujours comme des demi-cystines (paramètre E_ALL). Selon la seconde hypothèse, aucune cystéine n'apparaît comme une demi-cystine (paramètre E_NO). Des études ont montré que ce calcul est fiable pour les protéines qui contiennent du tryptophane. Par contre, il peut y avoir jusqu'à 10% d'erreur pour les protéines sans tryptophane.

- *GRAVY* : la valeur du paramètre GRAVY reflète le profil d'hydrophobicité/hydrophilie de la protéine. Ce paramètre a été défini comme le rapport de la somme des valeurs d'hydropathie des acides aminés et de la longueur de la protéine [Kyte et Doolittle, 1982] et calculé selon l'équation B.6 :

$$GRAVY = \frac{1}{L} * \sum_{i=1}^{L} H(aa_i) \tag{B.6}$$

où $H(aa_i)$ est la valeur hydropathique de l'acide aminé aa_i.

B.2.2 Caractérisation des propriétés des acides aminés

Nous avons considéré les 12 paramètres physico-chimiques suivants pour caractériser les acides aminés. Pour chaque paramètre, les acides aminés caractérisés par ce paramètre sont indiqués :

- polaire : R, N, D, E, Q, H, K, S, T, Y
- non polaire : A, C, G, I, L M, F, P, W, V
- aliphatique : I, V, L
- aromatique : F, W, Y, H
- hydrophobe : A, G, C, V, T, I, L, M, F, Y, W, H, K
- petit : P, N, A, G, C, S, T, V, D
- très petit : A, G, C, S

– chargé positivement (basique) : R, H, K
– chargé négativement (acide) : D, E

Nous avons calculé ces paramètres sur l'ensemble des acides aminés et normalisé par le nombre total d'acides aminés, c'est-à-dire la longueur de la protéine.

B.2.3 Caractérisation de la composition en acides aminés

Pour un acide aminé donné, comme la cystéine, nous avons calculé le nombre total de cet acide aminé dans la séquence de la protéine. Comme ce nombre dépend de la longueur de la protéine, nous avons voulu effectuer une normalisation pour éviter des déséquilibres provenant des différences de longueur et non des différences de composition en acides aminés. Pour cela, nous avons proposé deux normalisations : par rapport au nombre total d'atomes de la protéine ou par rapport à la longueur de la protéine (c'est l'option par défaut). De plus il est possible d'ajouter le clivage de la méthionine en position N-terminale qui signale le début de la traduction. Les règles implémentées sont les suivantes [Hirel *et al.*, 1989], [Moerschell *et al.*, 1990] :

– retirer la méthionine si elle est suivie par un acides aminés parmi glycine, alanine, cystéine, thréonine, proline, valine
– ne jamais retirer la méthionine si un acide aminé proline est touvé en troisième position

B.2.4 Caractérisation de la composition en atomes

Pour un atome donné, comme le soufre, nous avons calculé le nombre total d'atomes dans la séquence en acides aminés de la protéine. Comme le nombre d'atomes dépend de la longueur de la protéine, nous avons voulu effectuer une normalisation pour éviter des déséquilibres provenant des différences de longueur et non des différences de composition en atomes. Pour cela, nous avons proposé deux normalisations : par rapport au nombre total d'atomes de la protéine (c'est l'option par défaut) ou par rapport à la longueur de la protéine.

Annexe C

Tests statistiques

Nous présentons ici les différents tests statistiques utilisés au cours du travail, d'abord le test de Wilcoxon ayant permis de mettre en évidence des biais de composition entre deux groupes de protéines, puis le test hypergéométrique ayant permis d'évaluer la probabilité d'obtenir aléatoirement une intersection entre deux ensemble de interactions protéine-protéine.

C.1 Test de Wilcoxon

Le test de Wilcoxon, aussi appelé test des rangs, somme des rangs, ou test de Mann et Whitney, est un test statistique non paramétrique. Il est utilisé pour comparer deux populations à partir d'échantillons indépendants, dans le cas de données continues. Son principe est de classer l'ensemble des observations des deux échantillons (de taille n_1 et n_2) par ordre croissant, de déterminer les rangs de chacune d'entre elles dans cet ensemble, puis de calculer la somme des rangs relatifs, par exemple, au premier échantillon. Nous désignerons cette somme par la variable aléatoire X_1.

$$n_1 + n_2 \geq 30 \tag{C.1}$$

Pour des effectifs suffisamment élevés, c'est-à-dire vérifiant la condition C.1, la variable X_1 suit approximativement une loi normale de paramètres (μ, σ) calculés selon les équations C.2 et C.3.

$$\mu = \frac{n_1 * (n_1 + n_2 + 1)}{2} \tag{C.2}$$

$$\sigma = \sqrt{\frac{n_1 * n_2 * (n_1 + n_2 + 1)}{12}} \tag{C.3}$$

Par conséquent, la variable aléatoire Z_1 suit une loi normale centrée réduite (voir équation C.4.

$$Z_1 = \frac{X_1 - \mu}{\sigma} \tag{C.4}$$

Ainsi, pour un test bilatéral de niveau de probabilité α, on doit rejeter l'hypothèse d'identité des distributions des deux populations quand l'équation C.5 est vérifiée.

$$P(|Z_1| \geq z_1) \leq \alpha \tag{C.5}$$

C.2 La cinétique Cd chez *S. cerevisiae*

TAB. C.1: Résultats des tests statistiques sur les biais pour la cinétique Cd chez *S. cerevisiae*.

Biais	Normalisation	Wilcoxon	T-test
sulfur	**atom**	**0,00006**	**0,00008**
cysteine	**length**	**0,00013**	**0,00008**
polar	length	0,00039	0,01274
methionine	length	0,00075	0,00714
extraSmall	length	0,00865	0,09556
hydrophobic	length	0,00978	0,03262
GRAVY	none	0,01244	0,03641
negativelyCharged	length	0,01811	0,07145
molecular-weight	length	0,06592	0,07442
aspartate	length	0,06640	0,15153
isoelectric-point	none	0,07930	0,05843
charged	length	0,08737	0,12306
small	length	0,09870	0,27061
glycine	length	0,10273	0,08371
oxygen	atom	0,11264	0,20890
asparagine	length	0,11486	0,30446
lysine	length	0,12890	0,18693
leucine	length	0,17803	0,34686
positivelyCharged	length	0,21259	0,44956
instability-index	none	0,23095	0,18274
tyrosine	length	0,25309	0,12102
alanine	length	0,25444	0,18609
carbon	atom	0,27099	0,34892
valine	length	0,27524	0,23523
extinction-coefNO	none	0,29121	0,21492
serine	length	0,30322	0,34130
glutamine	length	0,32500	0,24123
aromatic	length	0,37677	0,27363
hydrogen	atom	0,38734	0,26707
tryptophane	length	0,38912	0,49594
arginine	length	0,40537	0,41910
threonine	length	0,42577	0,25532
glutamate	length	0,45648	0,28506
aliphatic	length	0,61351	0,62434
nitrogen	atom	0,64106	0,69640
phenylalanine	length	0,66440	0,56793
isoleucine	length	0,71448	0,42124
length	none	0,71931	0,53688
aliphatic-index	none	0,85613	0,55648
absorbanceNO	none	0,91235	0,97290

Suite sur la page suivante

Tab. C.1 – suite de la page précédente

Biais	Normalisation	Wilcoxon	T-test
absorbanceALL	none	0,93291	0,99632
proline	length	0,94837	0,81516
histidine	length	0,96901	0,78687

C.3 La cinétique Cd chez *Synechocystis*

C.3.1 Comparaison des gènes globalement régulés

Tab. C.2: Résultats des tests statistiques sur les biais pour la cinétique Cd chez *Synechocystis*.

Biais	Normalisation	Wilcoxon	T-test
hydrophobic	**length**	**0,00000**	**0,00000**
instability-index	**none**	**0,00000**	**0,00000**
glutamine	**length**	**0,00000**	**0,00000**
glycine	**length**	**0,00000**	**0,00000**
molecular-weight	**length**	**0,00000**	**0,00000**
small	**length**	**0,00001**	**0,00004**
valine	**length**	**0,00012**	**0,00003**
GRAVY	**none**	**0,00013**	**0,00005**
extraSmall	**length**	**0,00015**	**0,00003**
alanine	**length**	**0,00032**	**0,00011**
polar	**length**	**0,00036**	**0,00002**
hydrogen	atom	0,00115	0,00037
nitrogen	atom	0,00149	0,00394
glutamate	length	0,00320	0,00319
arginine	length	0,00343	0,03463
negativelyCharged	length	0,00384	0,00275
histidine	length	0,00515	0,00150
isoleucine	length	0,00892	0,01466
tryptophane	length	0,01263	0,04562
leucine	length	0,01321	0,02573
positivelyCharged	length	0,02963	0,29178
lysine	length	0,03159	0,01350
oxygen	atom	0,05344	0,04204
threonine	length	0,06061	0,07642
absorbanceALL	none	0,06773	0,15732
absorbanceNO	none	0,06977	0,16206
cysteine	length	0,09345	0,00946
isoelectric-point	none	0,09487	0,13720
charged	length	0,10124	0,06680
phenylalanine	length	0,12693	0,10416

Suite sur la page suivante

Tab. C.2 – suite de la page précédente

Biais	Normalisation	Wilcoxon	T-test
aliphatic-index	none	0,13495	0,05641
length	none	0,14640	0,39430
aliphatic	length	0,17885	0,08027
aspartate	length	0,17940	0,09586
aromatic	length	0,25566	0,43990
serine	length	0,32200	0,25507
tyrosine	length	0,36438	0,43623
methionine	length	0,40399	0,80312
asparagine	length	0,53668	0,26161
sulfur	atom	0,56546	0,20538
proline	length	0,57434	0,33699
extinction-coefNO	none	0,71654	0,59987
extinction-coefALL	none	0,72756	0,60380
carbon	atom	0,89081	0,69647

C.3.2 Comparaison des gènes répondant en deux phases

Tab. C.3: Résultats des tests statistiques sur les biais pour la cinétique Cd chez *Synechocystis*.

Biais	Normalisation	Wilcoxon	T-test
lysine	**length**	**0,00000**	**0,00000**
leucine	**length**	**0,00000**	**0,00000**
tryptophane	**length**	**0,00000**	**0,00000**
instability-index	**none**	**0,00000**	**0,00000**
glutamine	**length**	**0,00000**	**0,00002**
absorbanceALL	**none**	**0,00002**	**0,00001**
absorbanceNO	**none**	**0,00002**	**0,00002**
hydrophobic	**length**	**0,00006**	**0,00020**
alanine	length	0,00110	0,00012
small	length	0,00241	0,00359
isoleucine	length	0,00298	0,00143
molecular-weight	length	0,00324	0,00067
arginine	length	0,00433	0,02485
valine	length	0,00521	0,00193
nitrogen	atom	0,01358	0,06239
proline	length	0,01524	0,00633
methionine	length	0,02053	0,06367
carbon	atom	0,02317	0,03615
cysteine	length	0,02607	0,00620
charged	length	0,03143	0,11093
tyrosine	length	0,03282	0,04748

Suite sur la page suivante

Tab. C.3 – suite de la page précédente

Biais	Normalisation	Wilcoxon	T-test
aromatic	length	0,03771	0,05701
threonine	length	0,04176	0,03444
extinction-coefNO	none	0,04382	0,68291
hydrogen	atom	0,04406	0,03874
extinction-coefALL	none	0,04447	0,68361
oxygen	atom	0,04872	0,10211
glycine	length	0,05198	0,03684
extraSmall	length	0,05296	0,01414
histidine	length	0,11211	0,04738
aliphatic	length	0,13985	0,10714
aliphatic-index	none	0,19801	0,16096
length	none	0,20433	0,45231
positivelyCharged	length	0,21683	0,10926
serine	length	0,23536	0,18556
aspartate	length	0,24440	0,43944
negativelyCharged	length	0,27799	0,28259
glutamate	length	0,37853	0,35793
polar	length	0,46549	0,68782
GRAVY	none	0,61564	0,78072
isoelectric-point	none	0,67321	0,85262
phenylalanine	length	0,85109	0,96258
asparagine	length	0,89321	0,52519
sulfur	atom	0,90395	0,93711

C.4 La cinétique H_2O_2 chez *Synechocystis*

C.4.1 Comparaison des gènes globalement régulés

Tab. C.4: Résultats des tests statistiques sur les biais pour la cinétique ho chez *Synechocystis*.

Biais	Normalisation	Wilcoxon	T-test
alanine	length	0,02973	0,03154
polar	length	0,04005	0,05093
extinction-coefNO	none	0,07213	0,05563
extinction-coefALL	none	0,07436	0,05490
GRAVY	none	0,11506	0,11850
proline	length	0,12614	0,05724
length	none	0,15210	0,17596
isoleucine	length	0,15612	0,18817
serine	length	0,18193	0,20402
glutamate	length	0,21419	0,22237

Suite sur la page suivante

Tab. C.4 – suite de la page précédente

Biais	Normalisation	Wilcoxon	T-test
tyrosine	length	0,25810	0,32308
valine	length	0,26798	0,53631
hydrophobic	length	0,27812	0,15257
positivelyCharged	length	0,29489	0,23656
asparagine	length	0,30351	0,16154
aliphatic-index	none	0,32127	0,37609
glutamine	length	0,32582	0,58969
carbon	atom	0,34916	0,17044
lysine	length	0,35155	0,24511
nitrogen	atom	0,35395	0,27280
negativelyCharged	length	0,39893	0,28763
aliphatic	length	0,45275	0,55207
oxygen	atom	0,46116	0,21858
charged	length	0,49567	0,36050
aromatic	length	0,51642	0,50311
absorbanceALL	none	0,56548	0,41771
absorbanceNO	none	0,57177	0,42148
leucine	length	0,59724	0,74252
tryptophane	length	0,61017	0,63566
cysteine	length	0,62979	0,67877
molecular-weight	length	0,68337	0,38085
arginine	length	0,69019	0,47675
histidine	length	0,73860	0,83691
hydrogen	atom	0,78807	0,71773
phenylalanine	length	0,81314	0,90803
instability-index	none	0,81674	0,96472
aspartate	length	0,82395	0,80766
isoelectric-point	none	0,85292	0,94497
glycine	length	0,88942	0,78112
threonine	length	0,91511	0,98630
extraSmall	length	0,91878	0,52478
sulfur	atom	0,95565	0,91282
methionine	length	0,98891	0,64783

C.4.2 Comparaison des gènes répondant en deux phases

Tab. C.5: Résultats des tests statistiques sur les biais pour la cinétique H_2O_2 chez *Synechocystis*.

Biais	Normalisation	Wilcoxon	T-test
length	**none**	**0,00000**	**0,00000**
extinction-coefALL	**none**	**0,00000**	**0,00000**
Suite sur la page suivante			

Tab. C.5 – suite de la page précédente

Biais	Normalisation	Wilcoxon	T-test
extinction-coefNO	**none**	**0,00000**	**0,00000**
positivelyCharged	**length**	**0,00002**	**0,00012**
extraSmall	**length**	**0,00003**	**0,00030**
isoleucine	**length**	**0,00006**	**0,00021**
lysine	**length**	**0,00020**	**0,00014**
molecular-weight	length	0,00022	0,00224
small	length	0,00071	0,00578
charged	length	0,00177	0,00275
serine	length	0,00288	0,00675
histidine	length	0,00773	0,01267
alanine	length	0,01756	0,01287
aromatic	length	0,01830	0,01788
carbon	atom	0,04649	0,05786
oxygen	atom	0,05287	0,15996
tyrosine	length	0,06094	0,09506
proline	length	0,07470	0,28397
isoelectric-point	none	0,07638	0,17429
negativelyCharged	length	0,10667	0,12336
polar	length	0,10756	0,09143
absorbanceALL	none	0,11417	0,07468
glycine	length	0,11581	0,14140
absorbanceNO	none	0,11843	0,07466
aspartate	length	0,13726	0,14641
tryptophane	length	0,14383	0,12513
GRAVY	none	0,14835	0,07977
sulfur	atom	0,15982	0,02845
methionine	length	0,22230	0,02702
leucine	length	0,22697	0,12183
glutamine	length	0,25090	0,21881
glutamate	length	0,26330	0,27072
valine	length	0,26373	0,09101
phenylalanine	length	0,31263	0,69404
cysteine	length	0,32448	0,68597
arginine	length	0,37203	0,49753
asparagine	length	0,72699	0,79943
instability-index	none	0,76786	0,88574
threonine	length	0,77798	0,85496
hydrogen	atom	0,85303	0,70510
nitrogen	atom	0,89795	0,82681
hydrophobic	length	0,92297	0,91360
aliphatic	length	0,93105	0,73466
aliphatic-index	none	0,96428	0,59007

C.5 Test hypergéométrique

Pour évaluer la pertinence des recouvrements entre les interactions prédites et les interactions mises en évidence expérimentalement, nous avons calculé la probabilité d'obtenir par hasard un recouvrement au moins aussi grand que celui observé. Pour cela, nous avons utilisé un modèle hypergéométrique.

La distribution hypergéométrique est une distribution de probabilité discrète qui décrit le nombre de succès X au cours d'une séquence de S_2 tirages dans une population finie N sans replacement. Elle est décrite par les paramètres suivants :

- N = la taille de la population
- S_1 = le nombre de succès dans la population
- S_2 = la taille de l'échantillon
- X = le nombre de succès dans l'échantillon

Ainsi la probabilité d'obtenir exactement X succès est :

$$p(x = k) = \frac{\mathrm{C}_X^{S_1}\,\mathrm{C}_{S_2-X}^{N-S_1}}{\mathrm{C}_{S_2}^{N}} \tag{C.6}$$

D'où la probabilité d'en obtenir au moins X :

$$p(x >= k) = 1 - \sum_{k=0}^{X-1} p(k) \tag{C.7}$$

Les résultats du test hypergéométrique ont été calculés avec le logiciel R en utilisant la commande suivante :

$$phyper(X - 1, S_1, N - S_1, S_2, lower.tail = FALSE) \tag{C.8}$$

où

- S_1 est la taille du premier jeu de données, c'est-à-dire le nombre d'interactions identifiées expérimentalement,
- N est le nombre de paires de protéines de l'espace considéré. Quand l'information est disponible, N est calculé comme le produit des nombres de protéines appâts viables et proies viables ($N = VB \times VP$), c'est l'espace total des paires de protéines viables. Par défaut on considère que les 3505 protéines sont viables,
- S_2 est la taille du second jeu de données, c'est-à-dire le nombre d'interactions prédites *in-silico* appartenant à l'espace $VB \times VP$,
- X est le nombre d'interactions présents dans les deux jeux de données.

Les paramètres utilisés lors des différents calculs sont indiqués dans la table C.6.

Jeu 1	**Jeu 2**	**VB**	**VP**	**S1**	**S2**	**X**	**P-value**
InteroBH_HIGH	low-throughput exp	3 505	3 505	179	2 748	5	$7,8.10^{-10}$
InteroBH_MEDIUM	low-throughput exp	3 505	3 505	179	5 070	8	$1,8.10^{-14}$
InteroBH_LOW	low-throughput exp	3 505	3 505	179	8 586	10	$1,8.10^{-16}$
InteroPorc	low-throughput exp	3 505	3 505	179	1 446	2	$2,2.10^{-4}$
InteroFull	low-throughput exp	3 505	3 505	179	8 783	10	$2,2.10^{-4}$
InteroBH_HIGH	*SatoFull*	3 505	3 505	3 236	2 748	12	$2,1.10^{-11}$
InteroBH_MEDIUM	*SatoFull*	3 505	3 505	3 236	5 070	17	$1,0.10^{-13}$
InteroBH_LOW	*SatoFull*	3 505	3 505	3 236	8 586	24	$5,4.10^{-17}$
InteroPorc	*SatoFull*	3 505	3 505	3 236	1 446	6	$3,0.10^{-6}$
InteroFull	*SatoFull*	3 505	3 505	3 236	8 783	25	$8,1.10^{-18}$
SatoFull_Uni_Bi	All	976	1 259	2 990	3 904	25	$1,9.10^{-5}$

TAB. C.6 – **Paramètres du modèle hypergéométrique**

Annexe D

Classes de gènes

Nous présentons ici certaines classes de gènes et les profils d'expression correspondants. Dans un premier temps, il s'agit des gènes globalement régulés au cours de la cinétique Cd. Dans un second temps, il s'agit des gènes répondant en deux phases au cours de cette même cinétique.

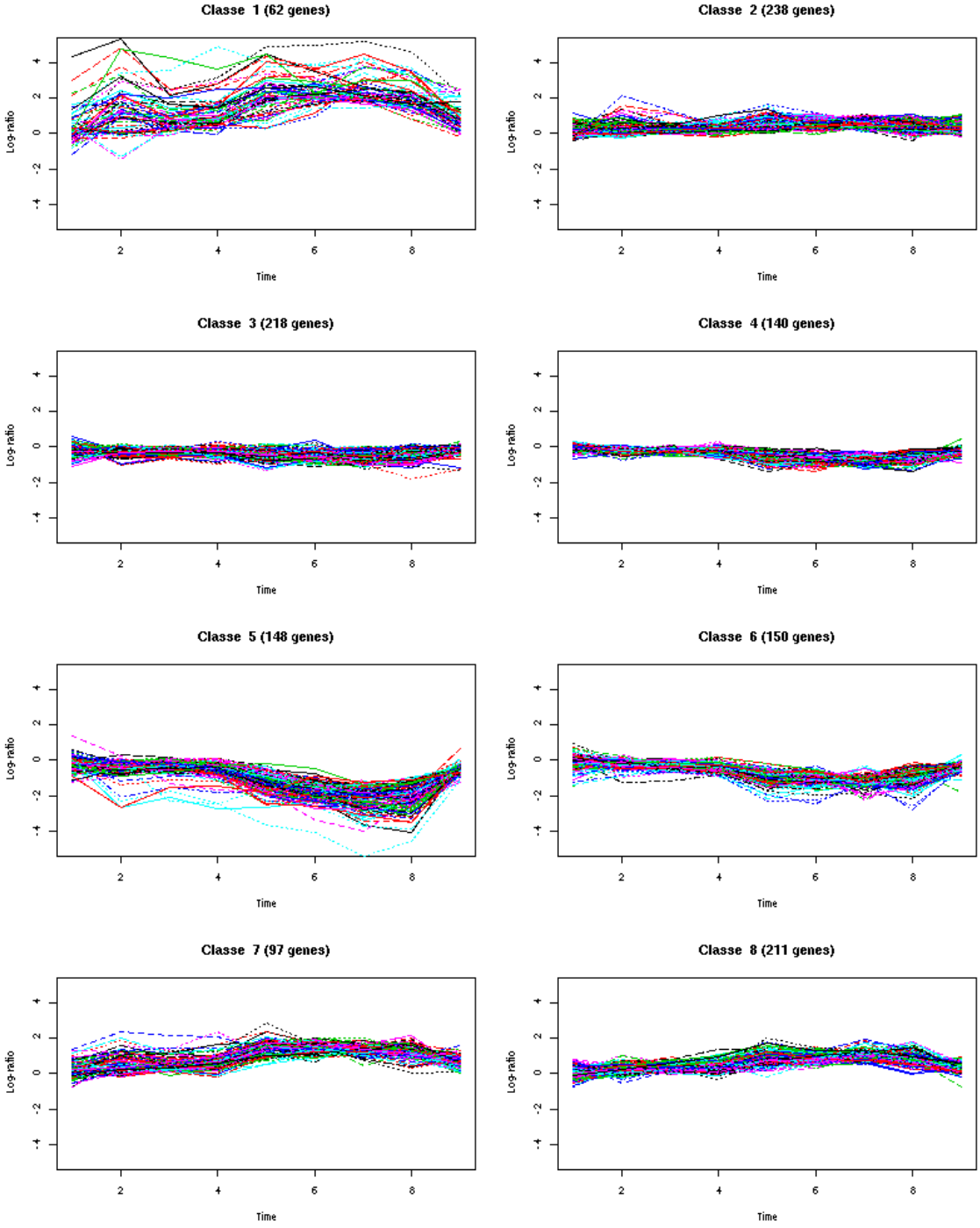


FIG. D.1 – **Classes de gènes globalement régulés pour le Cd.** Cette figure représente les huit classes identifiées dans l'ensemble des gènes globalement régulés au cours de la réponse au stress Cd. Les abscisses représentent les neuf temps de la cinétique.

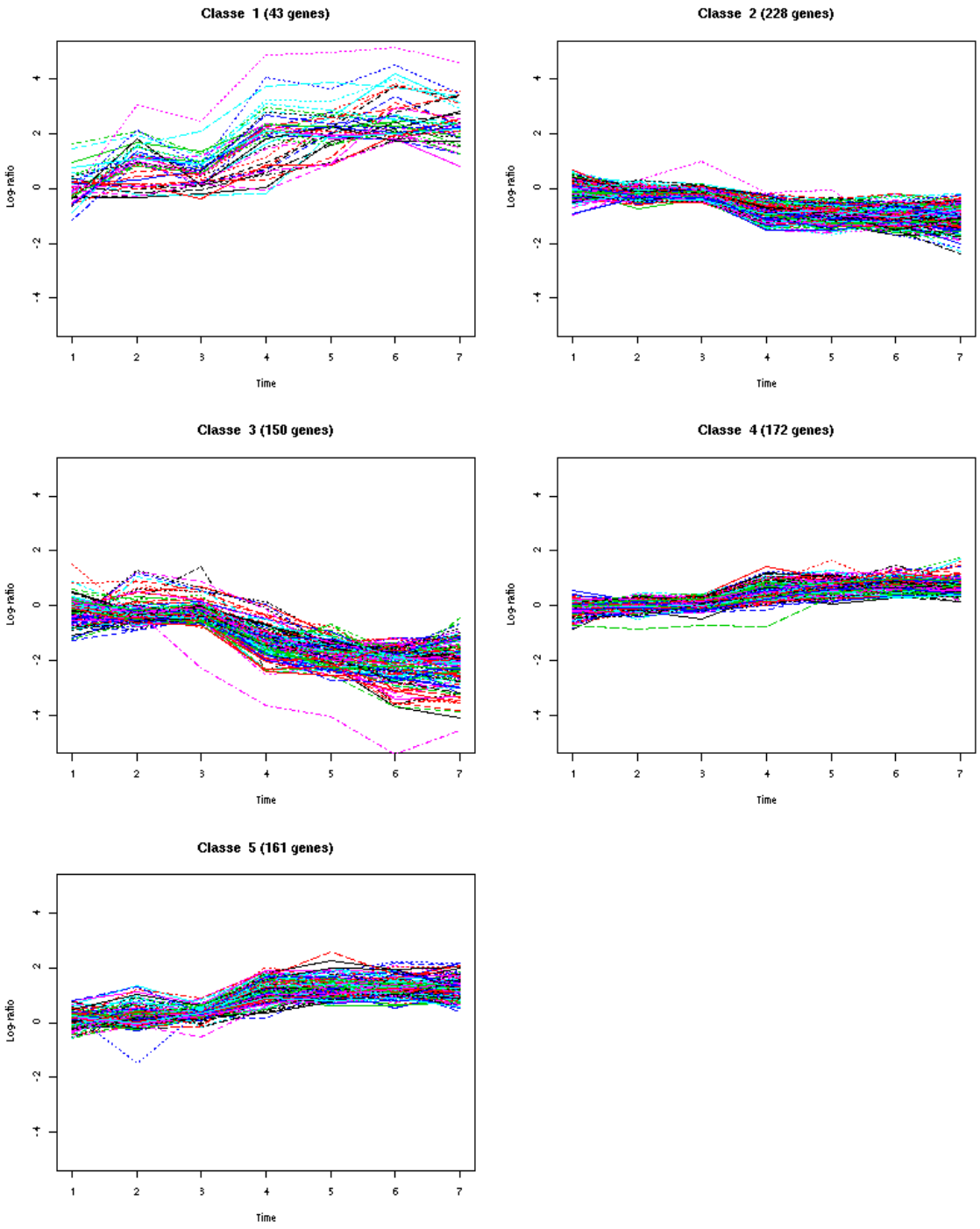


FIG. D.2 – **Classes de gènes répondant en deux phases pour le Cd.** Cette figure représente les différentes classes identifiées dans l'ensemble des gènes répondant en deux phases au cours de la réponse au stress Cd. Les abscisses représentent les différent temps de la cinétique constituant les deux phases de réponse (trois temps, puis quatre temps).

Annexe E

Mesure de similarité

La similarité fonctionnelle entre deux protéines a été quantifiée par une mesure de similarité. Cette mesure est définie comme la moyenne des mesures de similarité entre les ensembles de termes GO associés à chacune des deux protéines [Lubovac *et al.*, 2006] :

$$ss(p_k, p_l) = \frac{1}{m \times n} \sum_{t_i \in T_i, t_j \in T_j} sim(t_i, t_j) \tag{E.1}$$

où

- T_k est l'ensemble des m termes de l'ontologie GO associés à la protéine p_k
- T_l est l'ensemble des n termes de l'ontologie GO associés à la protéine p_l
- $sim(t_i, t_j)$ est la mesure de similarité terme à terme définie par Lin *et al.* [Lin, 1998] et rappelée dans l'Équation E.2

$$sim(t_i, t_j) = \frac{2ln\{p_{ms}(t_i, t_j)\}}{ln\{p(t_i)\} + ln\{p(t_j)\}} \tag{E.2}$$

où

- $p(t_i)$ est la probabilité du terme t_i
- $p_{ms}(t_i, t_j)$ est la probabilité du plus petit "subsumer" de t_i et t_j. Cette probabilité est définie comme la probabilité la plus faible trouvée parmi les termes parents partagés par t_i et t_j

Annexe F

Informations sur l'outil InteroPorc

On trouve ici le fichier qui donne toutes les informations nécessaires à l'utilisation autonome de l'outil *InteroPorc* afin de prédire des interactions protéine-protéine. Ce fichier est disponible avec le code source. Certaines informations sont également accessibles directement depuis l'interface web http ://biodev.extra.cea.fr/interoporc.

```
################################

  INTEROPORC PREDICTION MODULE

################################

http://biodev.extra.cea.fr/interoporc/

Magali Michaut
michaut.bioinfo@gmail.com
magali.michaut.2005@ingenieurs-supelec.org
mmichaut@ebi.ac.uk
--
07/06/20 created
07/11/26 added a simple run for one species predictions
07/11/30 added a result file with all source interactions used in the process
08/01/25 structured this file (content) and added the new project page
08/03/27 added PSI25-XML result files info
08/04/04 added info on new options to run the tool
--

You can find information on EBI website: http://www.ebi.ac.uk/~mmichaut/
and a document on this prediction programme: http://www.ebi.ac.uk/~mmichaut/documents/bio.pdf
but it is not updated any longer... ;-)
otherwise try http://people.rez-gif.supelec.fr/mmichaut/


******************************
CONTENT
******************************
1. How to use the module?
2. What are the result files?
3. Log4j property file example
4. FAQ
5. License
******************************


1) HOW TO USE THE MODULE?
=========================

You can either use the module online (see I)
or use the independant JAR file available on the project page (http://biodev.extra.cea.fr/interoporc/) (see II)
or import the latest jar library and include it into your code to use more options (see III).

>>> I) Online
Go to http://biodev.extra.cea.fr/interoporc/ and run analysis for the NCBI taxid you are interested in.
See all species of Integr8 on
http://www.ebi.ac.uk/integr8/OrganismSearch.do?action=setOrganismSearchType&searchType=2&pageContext=207

>>> II) With an independant JAR file (including all dependancies)
This JAR file is downloadable from the project page http://biodev.extra.cea.fr/interoporc/data/interopor.tar.gz
Here is described a simple way to use this program to predict interactions for one species.

If you have a jar with all dependencies --> interoporc.jar:
1) create a directory for the predictions --> DIR
2) put the jar in it
3) put a MITAB25 file in it with all interactions you want to use as source interactions from other species --> sourceInteractions.mitab
(if you're asking what the MITAB25 format could be, see the FAQ at the end)
4) download the orthologous clusters from ftp://ftp.ebi.ac.uk/pub/databases/integr8/porc/proc_gene.dat and put it in the directory
--> porc_gene.dat
```

```
5) choose the NCBI taxid of the species you are interested in
(for example Synechocystis is 1 148, yeast is 4 932, E. coli is 562 ... see all species of Integr8 on
http://www.ebi.ac.uk/integr8/OrganismSearch.do?action=setOrganismSearchType&searchType=2&pageContext=207)
6) OPTION: you can put a log4j-property-file in the dir  --> user.interoporc.log4j.properties
(you can copy-paste the example given below and put it in user.interoporc.log4j.properties file in the directory)

Then, execute this command in the directory DIR with your taxid (instead of 1 148):
Then you can use the tool with the main following options:
usage: Interoporc [OPTIONS]
Options:
usage: Interoporc [OPTIONS]
Options:
 -o,--output-directory <file>   Directory where all files will be created
 -i,--mitab-file <file>         MITAB File (Release 2.5) with source
                                interactions
 -n,--node-file <file>          NCBI Taxonomy file with taxonomy nodes
 -p,--porc-file <file>          PORC file with orthologous clusters
 -x,--xml-files                 If output XML files are required
 -c,--check-taxid               If protein accession numbers and taxids
                                are checked between interaction and porc data
 -m,--max-nb-inter-xml <int>    Maximum nb of interactions to generate a
                                XML file
 -h,--help                      print this message
 -l <file>                      use given file for log
 -t,--taxid <int>               NCBI taxonomy identifier of the species

 The required options are -i (input source interaction file) and -t (taxid).
 The option -l is highly recommended (log4j configuration file) otherwise you won't have any message.
 Here are some examples:
* To print options
java -cp interoporc.jar uk.ac.ebi.intact.interolog.prediction.RunForOneSpecies
java -cp interoporc.jar uk.ac.ebi.intact.interolog.prediction.RunForOneSpecies -h

* To predict interactions for Synechocystis (taxid=1 148)
java -ms500m -mx1200m -cp interoporc.jar uk.ac.ebi.intact.interolog.prediction.RunForOneSpecies -i sourceInteractions.mitab -t 1148
-l user.interoporc.log4j.properties

java -ms500m -mx1200m -cp interoporc.jar uk.ac.ebi.intact.interolog.prediction.RunForOneSpecies -o . -i sourceInteractions.mitab
-p porc_gene.dat -t 1148 -l user.interoporc.log4j.properties


>>> III) With the JAR available on EBI maven repos
http://www.ebi.ac.uk/~maven/m2repo/uk/ac/ebi/intact/dataexchange/psimi/intact-interolog-prediction/
http://www.ebi.ac.uk/~maven/m2repo_snapshots/uk/ac/ebi/intact/dataexchange/psimi/intact-interolog-prediction/2.0.0-SNAPSHOT/
You have to create an instance of InterologPrediction with the required working directory (where files will be created).
Then you can change some parameters if needed and finally just run it. An example is given below:

  InterologPrediction p = new InterologPrediction(workingDir);
  p.setClog(clogFile);
  p.setMitab(mitabFile);
  p.setPredictedinteractionsFileExtension(".mitab");
  p.setWriteDownCastHistory(true);

  p.setDownCastOnAllPresentSpecies(false);
  p.setClassicPorcFormat(false);
  p.setWriteDownCastHistory(true);
  p.setWriteSrcInteractions(true);
  ClogInteraction.setNB_LINES_MAX(100 000);
  p.setWritePorcInteractions(false);
  p.setDownCastOnChildren(false);

  p.run();

Be aware that this program needs some space. I am used to running it with extended arguments to the VM (-ms500m -mx1200m).
On the other hand, it does not take too much time.
Running it on the global MITAB25 file (merge of all Intact, MINT and DIP)
and predicting interactions for all species present in it will take less than 5 minutes.



2) WHAT ARE THE RESULT FILES?
=============================
1. InteroPorc.predictedInteractions.mitab / InteroPorc.predictedInteractions.xml
Predicted interactions are described in both PSIMI25-XML and MITAB25 formats
(PSI25-XML is obtained with option -x and if not too many interactions are predicted)

2. KnownInteractions.mitab / KnownInteractions.xml
Interactions extracted from the source interaction fils for the species you are interested in
are described in both PSIMI25-XML and MITAB25 formats.
(PSI25-XML is obtained with option -x and if not too many interactions are predicted)

3. AllInteractions.mitab / AllInteractions.xml
All interactions (known and predicted) are described in both PSIMI25-XML and MITAB25 formats.
(PSI25-XML is obtained with option -x and if not too many interactions are predicted)

4. srcInteractionsUsed.txt
All source interactions used during the process are described in the srcInteractionsUsed.txt file.

5. interologPrediction.log
Comments are written in the interologPrediction.log file during the process if you have configured the log4j property file.

6. downCast.history.txt
Some information about the constructed porc interactions are in the tabulated text file downCast.history.txt
porcA=id from the porc data
porcB=id from the porc data
prot1=number of proteins in cluster porcA
prot2=number of proteins in cluster porcB
sources=number of source interactions used to construct this porc interaction
```

```
inferences=number of interaction predicted thanks to this porc interaction

0. proteome_report.txt
The proteome_report.txt file is downloaded and used during the process. It is not a result file but rather an input file.
You can remove it or keep it in the directory so that it is not downloaded again next time.


If it is not clear, don't hesitate to contact me.
Have fun! :-)



3) LOG4J PROPERTY FILE EXAMPLE
==============================

log4j.rootCategory=INFO, R, A

# package/class specific config
log4j.category.psidev=ERROR

# ***** A is set to be a ConsoleAppender.
log4j.appender.A=org.apache.log4j.ConsoleAppender
log4j.appender.A.layout=org.apache.log4j.PatternLayout
log4j.appender.A.layout.ConversionPattern=\%m\%n
log4j.appender.A.Threshold=WARN

# ***** R file appender
log4j.appender.R=org.apache.log4j.RollingFileAppender
log4j.appender.R.File=interoporc.log
log4j.appender.R.MaxFileSize=1000KB
log4j.appender.R.MaxBackupIndex=0
log4j.appender.R.layout=org.apache.log4j.PatternLayout
log4j.appender.R.layout.ConversionPattern=\%d - \%m\%n



4) FAQ
========

* What is the PSI25-XML format?
PSI25-XML is the standard molecular interaction data exchange format defined by the Proteomics Standards Initiative (PSI).
All information are on the PSI website: http://www.psidev.info/

* What is the MITAB25 format?
MITAB25 describes binary interactions, one pair of interactors per row. Columns are separated by tabulators.
Fore more information, see:
- a simple readme file ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/README
- the Proteomics Standards Initiative (PSI) website http://www.psidev.info/


5) LICENSE
===========
Copyright (c) 2002 The European Bioinformatics Institute, and others.
All rights reserved.

 Redistribution and use in source and binary forms, with or without
modification, are permitted provided that the following conditions
are met:

1. Redistributions of source code must retain the above copyright
notice, this list of conditions and the following disclaimer.

2. Redistributions in binary form must reproduce the above copyright
 notice, this list of conditions and the following disclaimer in
the documentation and/or other materials provided with the
distribution.

3. The name IntAct must not be used to endorse or promote products
 derived from this software without prior written permission. For
written permission, please contact intact-dev@ebi.ac.uk

4. Products derived from this software may not be called "IntAct"
nor may "IntAct" appear in their names without prior written
permission of the IntAct developers.

5. Redistributions of any form whatsoever must retain the following
acknowledgment:
 "This product includes software developed by IntAct
 (http://www.ebi.ac.uk/intact)"

THIS SOFTWARE IS PROVIDED BY THE INTACT GROUP ``AS IS'' AND ANY
EXPRESSED OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE
IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR
PURPOSE ARE DISCLAIMED.  IN NO EVENT SHALL THE INTACT GROUP OR
ITS CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL,
SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT
NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES;
LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION)
HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT,
STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE)
ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED
OF THE POSSIBILITY OF SUCH DAMAGE.
```

Annexe G

Jeux de données d'interactions protéine-protéine

Dans le cadre de l'étude des interactions protéine-protéine mises en évidence chez *Synechocystis* par une approche de double-hybride [Sato *et al.*, 2007], nous avons considéré les jeux de données majeurs d'interactions protéine-protéine publiés jusqu'à présent chez la levure et *H. pylori* grâce à des identification expérimentales par *Y2H* ou *AP-MS*.

G.1 Obtention des données

Les interactions mises en évidences par Sato *et al.* ont été obtenues directement par la publication [Sato *et al.*, 2007]. Les interactions mises en évidence par Rain *et al.* ont été obtenues par une requête sur la publication (pubmed id = 11196647) dans la base de données MINT [Zanzoni *et al.*, 2002]. Les interactions mises en évidence chez la levure ont été obtenues grâce au package *ppiData* du logiciel R/Bioconductor [Chiang *et al.*, 2007].

G.2 Caractérisation des données

Nom	**Année**	**Organisme**	**Techno**	**PPI**	**Prot**	$\overline{\mathbf{CC}}$	$\overline{\mathbf{k}}$**/N**
Sato	2007	*Synechocystis*	Y2H	3 236	1 920	0,07	0,0017
Krogan	2006	yeast	AP-MS	61 391	5 361	0,40	0,004
Gavin	2006	yeast	AP-MS	18 028	2 551	0,22	0,006
Zhao	2005	yeast	Y2H	90	91	0,00	0,022
Krogan	2004	yeast	AP-MS	1 043	485	0,76	0,008
Hazbun	2003	yeast	Y2H	2 520	1 978	0,09	0,001
Ho	2002	yeast	AP-MS	3 618	1 578	0,03	0,003
Gavin	2002	yeast	AP-MS	3 226	1 362	0,15	0,003
Tong	2002	yeast	AP-MS	61 391	5 361	0,40	0,042
Rain	2001	*H. pylori*	Y2H	1 568	740	0,14	0,005
Ito	2001	yeast	Y2H	4 449	3 242	0,11	0,002
Cagney	2001	yeast	Y2H	51	48	0,10	0,044
Uetz-1	2000	yeast	Y2H	942	996	0,07	0,002
Uetz-2	2000	yeast	Y2H	517	503	0,06	0,004

TAB. G.1 – Liste des réseaux d'interactions protéine-protéine. Pour chaque jeu de données sont indiqués le nom, l'année de publication des données, l'organisme d'étude, l'approche expérimentale utilisée, le nombre d'interactions protéine-protéine, le nombre de protéines en interaction, le coefficient de clustering moyen et le rapport entre le degré moyen et le nombre de nœuds.

Annexe H

Topologie

H.1 Étude des données expérimentales

Tous les calculs permettant d'identifier les protéines appât et proie viables ont été faits avec le package *ppiStats* (version 1.4.1) dans R/Bioconductor [Gentleman *et al.*, 2004]. Les scores ont été calculés selon l'équation H.1.

$$z = \frac{n_{in} - n_{out}}{\sqrt{n_{in} + n_{out}}} \tag{H.1}$$

où n_{in} est le degré entrant non symétrique, c'est-à-dire le nombre d'appâts ayant détecté la protéine mais n'ayant pas été détectés par elle, et n_{out} est le degré sortant non symétrique, c'est-à-dire le nombre de proies détectées n'ayant pas détecté la protéine.

En se basant sur le modèle binomial du package *ppiStats*, 13 protéines VBP ont été identifiées avec un biais systématique. Nous avons utilisé un seuil de 10^{-2} sur la p-value pour sélectionner ces 13 protéines sont les suivantes : sll0149, sll0252, sll0750, sll0861, sll0985, sll1614, sll2012, slr0280, slr0798, slr0992, slr1198, slr1644 et slr2098.

H.2 Définition des paramètres topologiques

H.2.1 Définitions des paramètres

- Un **graphe** $G(N, M)$ consiste en un ensemble de N nœuds reliés entre eux par M arêtes. L'arête m_{ij} connecte les nœuds n_i et n_j.
- Le **voisinage** V_i d'un nœud i est l'ensemble des voisins directement connectés $V_i = \{n_j | m_{ij} \in M\}$.
- Le **degré** k_i d'un nœud n_i est le nombre d'arêtes incidentes.
- Le **diamètre** d'un graphe est le plus long plus court chemin entre deux nœuds.
- Le **coefficient de clustering** CC_i mesure la proportion de voisins connectés pour un nœud donné n_i.

$$CC_i = \frac{2|\{m_{jk}\}|}{k_i(k_i - 1)}, n_j \in N_i, n_k \in N_i \tag{H.2}$$

- Le **coefficient de voisinage** ou coefficient topologique quantifie dans quelle mesure un nœud partage des voisins avec les autres nœuds. Le coefficient topologique TC_i du nœud $_i$ possédant k_i voisins est défini selon l'équation H.3

$$TC_i = \frac{moy_j J(i,j)}{k_i} \tag{H.3}$$

où $J(i,j)$ est défini pour tous les nœuds j qui partagent au moins un voisin avec n_i. La valeur de $J(i,j)$ est le nombre de voisins partagés par les nœuds n_i et n_j plus un s'ils sont reliés par un arc.

H.2.2 Comparaisons des paramètres

- **GDD** : est la **distribution des degrés des graphlets**. La similarité entre les paramètres GDD de deux graphes est calculée selon l'approche proposée par Przulj [Przulj, 2007], en se basant soit sur une moyenne arithmétique (**GDD_A**), soit sur une moyenne géométrique (**GDD_G**). Ce paramètre prend des valeurs entre 0 et 1, où une valeur de 1 signifie que les graphes sont identiques pour cette propriété.
- **RFG** : indique la **fréquence relative des graphlets**. Une distance est calculée pour comparer les RFG de deux graphes [Przulj *et al.*, 2004].

H.3 Simulations avec des modèles de graphes

Les simulations ont été réalisées avec le logiciel GraphCrunch [Milenkovic *et al.*, 2008]. Pour un graphe réel de N nœuds et M arêtes, chaque graphe aléatoire généré contient N nœuds et M arêtes à 1% près.

Le modèle *er* est obtenu par un choix aléatoire de M paires parmi les N noeuds.

Le modèle *er_dd* conserve la distribution des degrés du graphe réel. Pour cela, une séquence de degrés est calculée, correspondant aux degrés des N nœuds. Un nœud de degré k se voit attribuer un paquet de taille k. Des paires de paquets sont tirées aléatoirement afin de créer les arêtes entre les nœuds correspondants, la taille des paquets étant ensuite réduite de 1 pour les deux paquets considérés.

Le modèle *geo* est obtenu en positionnant N points dans un espace métrique aléatoirement et en reliant ceux qui sont suffisamment proches l'un de l'autre (les M premières paires).

Le modèle *sf* est obtenu par la croissance avec attachement préférentiel [Barabasi et Albert, 1999]. On part d'un graphe vide et on ajoute progressivement N nœuds. À chaque étape, le nouveau nœud est connecté avec un nombre fixé de nœuds déjà présents dans le graphe avec une probabilité d'autant plus grande que le degré du nœud est élevé (équation H.4).

$$p(k_i) = \frac{k_i}{\sum_{j=1}^{N} k_j} \tag{H.4}$$

Le modèle *sticky* est obtenu en calculant pour chacun des N nœuds un indice dit de *stickiness* qui représente la tendance qu'a un nœud à être lié à d'autres. L'idée est ici de représenter la présence de domaines de liaisons des protéines au sein des réseaux d'interactions protéine-protéine. L'indice d'un nœud i est défini par l'équation H.5. Une arête est créée entre les nœuds i et j avec la probabilité $\theta_i\theta_j$.

$$\theta_i = \frac{k_i}{\sum_{j=1}^{N} k_j} \tag{H.5}$$

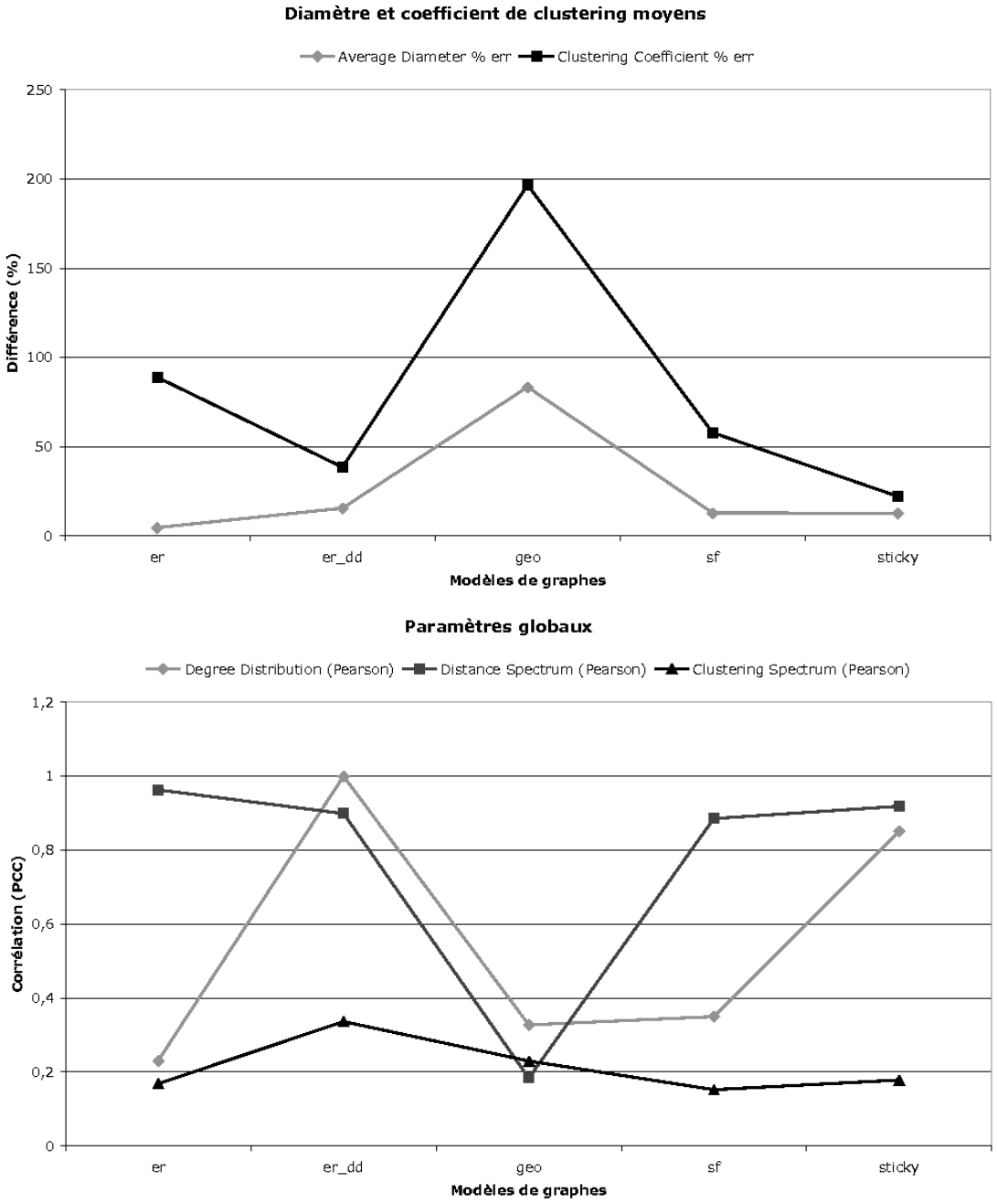


FIG. H.1 – **Comparaison des paramètres globaux pour *InteroPorc*.** Cette figure illustre la comparaison des paramètres globaux du réseau réel *InteroPorc* et de 50 instances de chacun des modèles aléatoires (*er*, *er_dd*, *geo*, *sf*, *sticky*). Le graphe du haut montre la différence en pourcentage entre les diamètres moyens, et également entre les coefficients de clustering moyens. Le graphe du bas montre les coefficients de corrélation entre les distributions des degrés, les distributions des plus courts chemins et les distributions des coefficients de clustering.

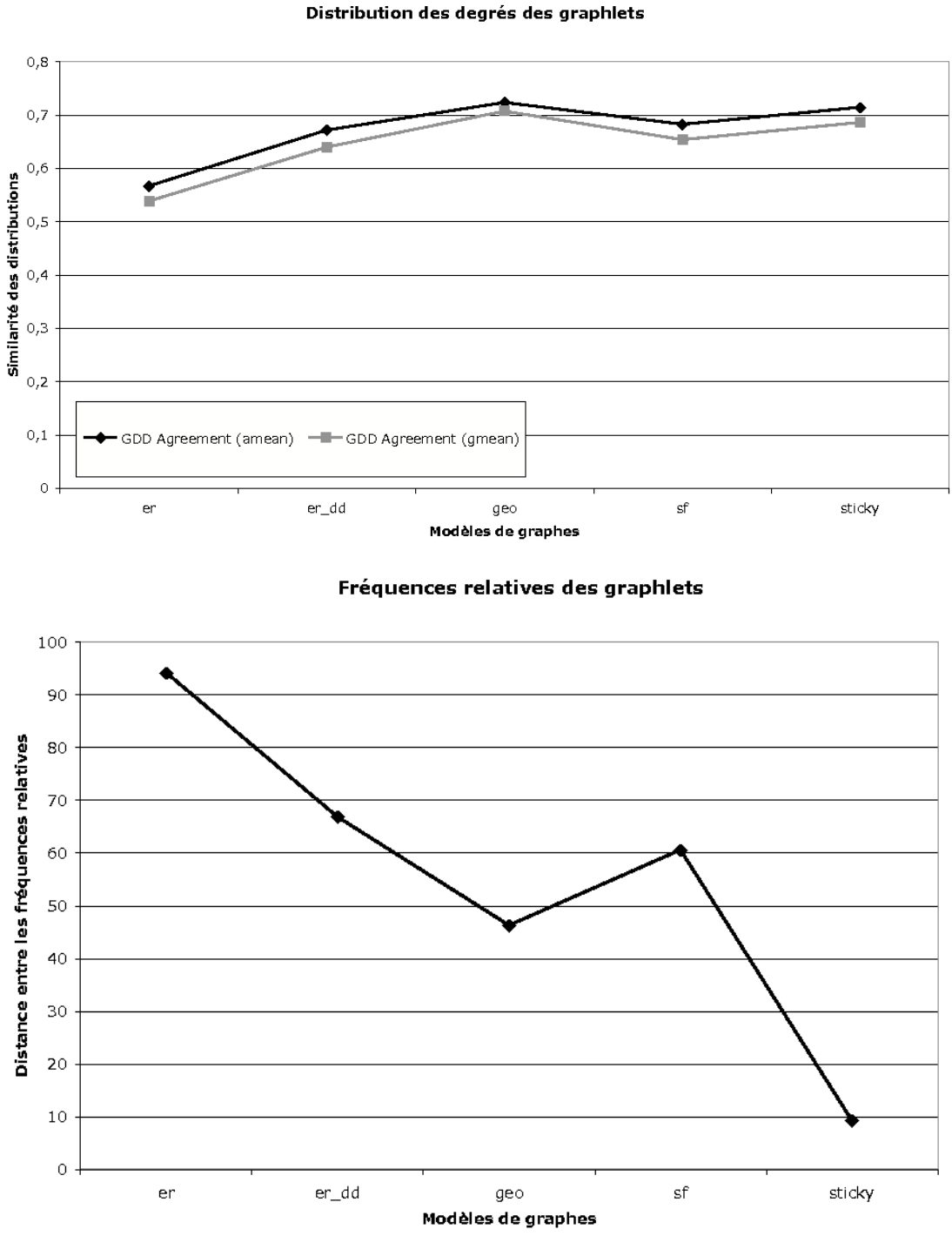


FIG. H.2 – **Comparaison des paramètres locaux pour *InteroPorc*.** Cette figure illustre la comparaison des paramètres locaux du réseau réel *InteroPorc* et de 50 instances de chacun des modèles aléatoires (*er*, *er_dd*, *geo*, *sf*, *sticky*). Le graphe du haut montre la similarité entre les distributions des degrés des graphlets. Le graphe du bas montre la distance entre les fréquences relatives des graphlets.

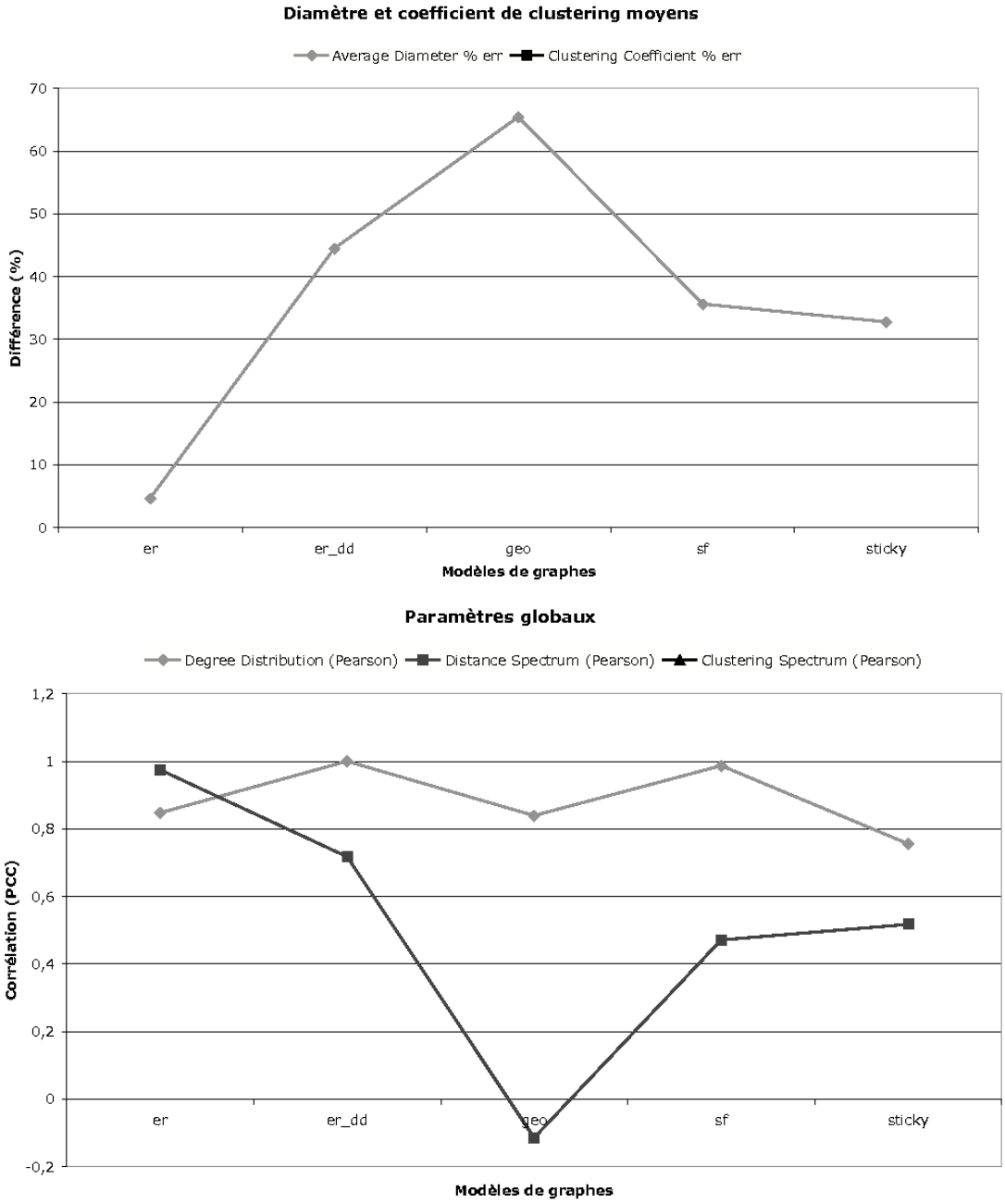


FIG. H.3 – **Comparaison des paramètres globaux pour *SatoCore*.** Cette figure illustre la comparaison des paramètres globaux du réseau réel *SatoCore* et de 50 instances de chacun des modèles aléatoires (*er*, *er_ dd*, *geo*, *sf*, *sticky*). Le graphe du haut montre la différence en pourcentage entre les diamètres moyens, et également entre les coefficients de clustering moyens. Le graphe du bas montre les coefficients de corrélation entre les distributions des degrés, les distributions des plus courts chemins et les distributions des coefficients de clustering.

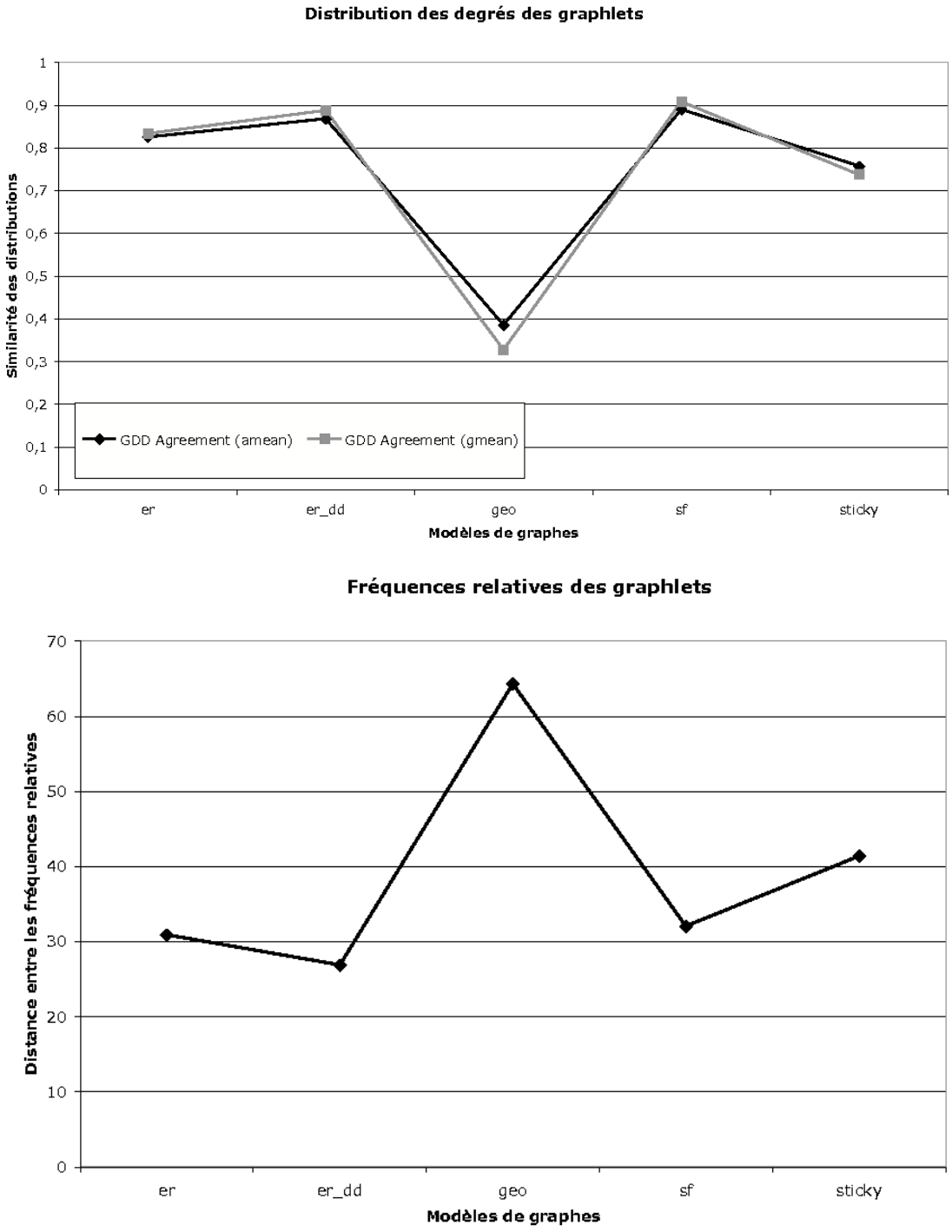


FIG. H.4 – **Comparaison des paramètres locaux pour *SatoCore*.** Cette figure illustre la comparaison des paramètres locaux du réseau réel *SatoCore* et de 50 instances de chacun des modèles aléatoires (*er*, *er_ dd*, *geo*, *sf*, *sticky*). Le graphe du haut montre la similarité entre les distributions des degrés des graphlets. Le graphe du bas montre la distance entre les fréquences relatives des graphlets.

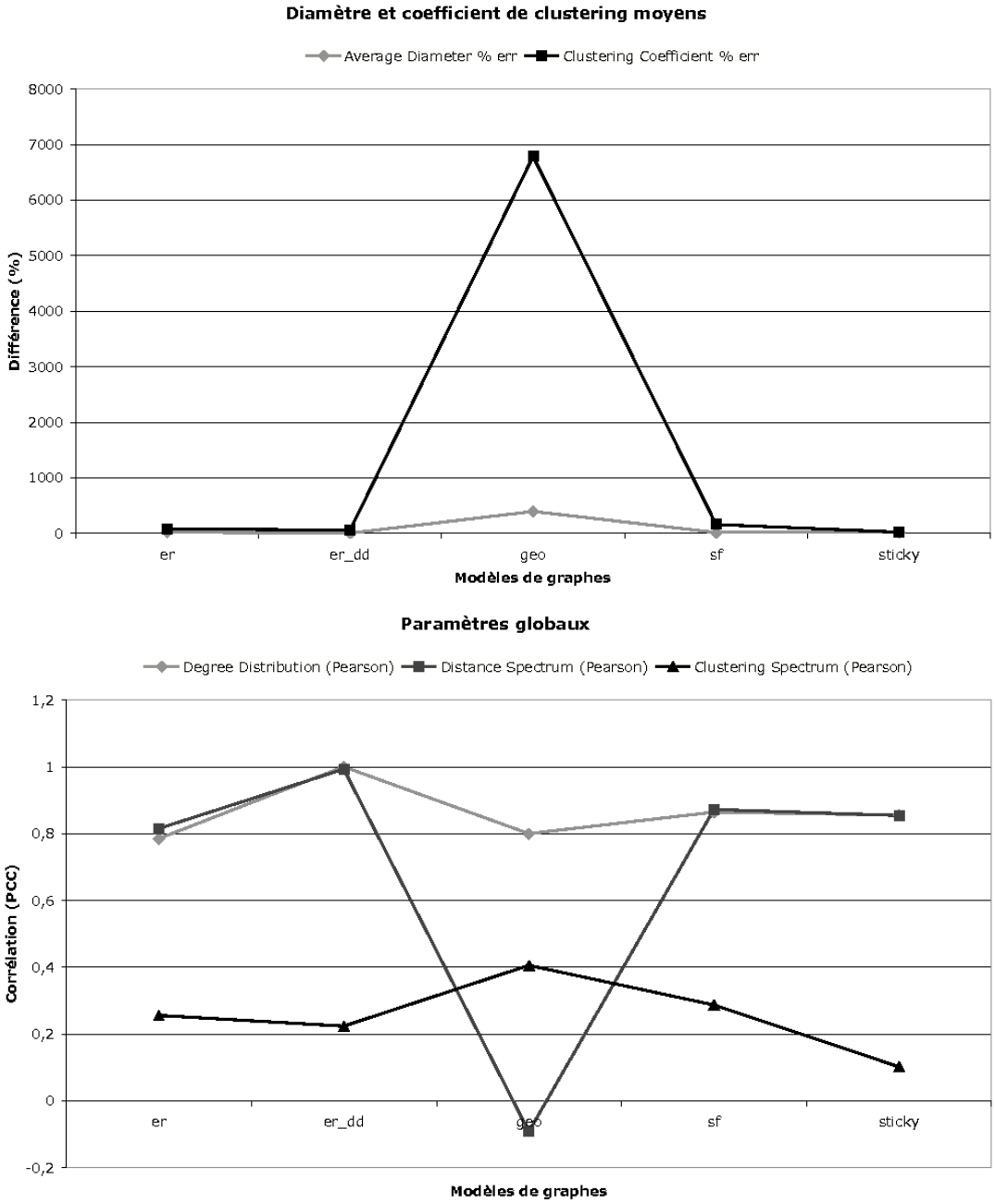


FIG. H.5 – **Comparaison des paramètres globaux pour *SatoFull*.** Cette figure illustre la comparaison des paramètres globaux du réseau réel *SatoFull* et de 50 instances de chacun des modèles aléatoires (*er*, *er_ dd*, *geo*, *sf*, *sticky*). Le graphe du haut montre la différence en pourcentage entre les diamètres moyens, et également entre les coefficients de clustering moyens. Le graphe du bas montre les coefficients de corrélation entre les distributions des degrés, les distributions des plus courts chemins et les distributions des coefficients de clustering.

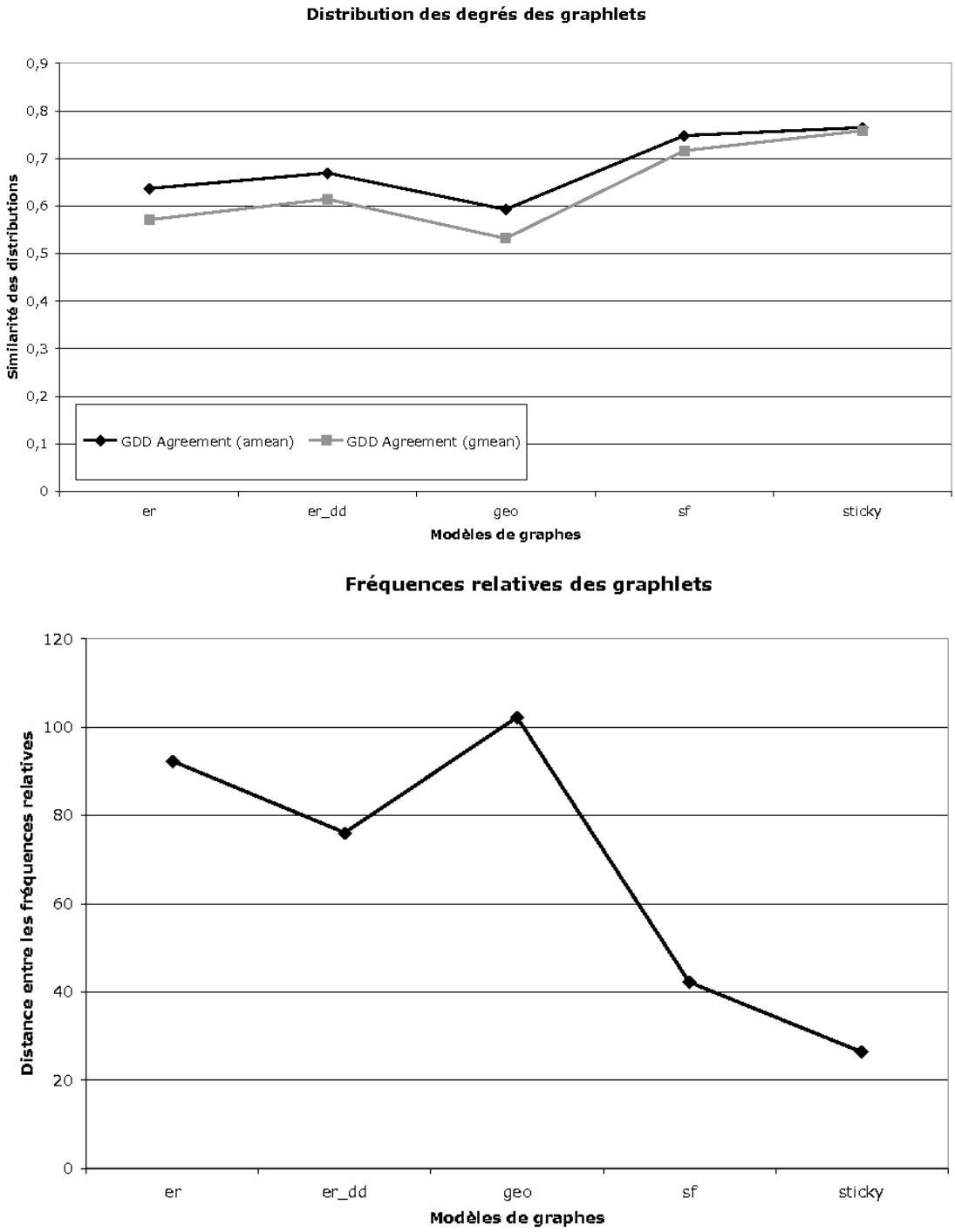


FIG. H.6 – **Comparaison des paramètres locaux pour *SatoFull*.** Cette figure illustre la comparaison des paramètres locaux du réseau réel *SatoFull* et de 50 instances de chacun des modèles aléatoires (*er*, *er_dd*, *geo*, *sf*, *sticky*). Le graphe du haut montre la similarité entre les distributions des degrés des graphlets. Le graphe du bas montre la distance entre les fréquences relatives des graphlets.

Annexe I

Identification de modules

I.1 Décomposition des temps des cinétiques par MCL

Ces classes **C**, de taille moyenne **Moy**, ont été identifiées à l'aide de l'algorithme MCL [Enright *et al.*, 2002]. Les interactions ont été pondérées pour chaque temps **Tps** de la cinétique **Stress** par la similarité des taux d'expression des gènes correspondants aux deux protéines en interaction. La colonne **Réseau** indique le réseau d'interactions protéine-protéine considéré. Les colonnes **Prot** et **PPI** indiquent respectivement le nombre de protéines et d'interactions du réseau considéré. La colonne **Max** indique la taille de la classe la plus grande.

Réseau	**Stress**	**Tps**	**Prot**	**PPI**	**C**	**Max**	**Moy**
InteroFull	Cd	1	1 011	8 783	306	37	3,30
InteroFull	Cd	2	1 011	8 783	304	35	3,33
InteroFull	Cd	3	1 011	8 783	267	49	3,79
InteroFull	Cd	4	1 011	8 783	297	33	3,40
InteroFull	Cd	5	1 011	8 783	302	65	3,35
InteroFull	Cd	6	1 011	8 783	310	40	3,26
InteroFull	Cd	7	1 011	8 783	304	41	3,33
InteroFull	Cd	8	1 011	8 783	301	33	3,36
InteroFull	Cd	9	1 011	8 783	297	34	3,40
InteroFull	H_2O_2	1	1 011	8 783	265	92	3,82
InteroFull	H_2O_2	2	1 011	8 783	277	100	3,65
InteroFull	H_2O_2	3	1 011	8 783	285	51	3,55
InteroFull	H_2O_2	4	1 011	8 783	275	89	3,68
InteroFull	H_2O_2	5	1 011	8 783	288	35	3,51
SatoFull	Cd	1	1 920	3 213	623	41	3,08
SatoFull	Cd	2	1 920	3 213	613	26	3,13
SatoFull	Cd	3	1 920	3 213	616	33	3,12
SatoFull	Cd	4	1 920	3 213	625	22	3,07
SatoFull	Cd	5	1 920	3 213	623	34	3,08
SatoFull	Cd	6	1 920	3 213	623	42	3,08
SatoFull	Cd	7	1 920	3 213	624	39	3,08
SatoFull	Cd	8	1 920	3 213	617	35	3,11

Suite sur la page suivante

Tab. I.1 – suite de la page précédente

Réseau	Stress	Tps	Prot	PPI	C	Max	Moy
SatoFull	Cd	9	1 920	3 213	609	34	3,15
SatoFull	H_2O_2	1	1 920	3 213	557	34	3,45
SatoFull	H_2O_2	2	1 920	3 213	561	33	3,42
SatoFull	H_2O_2	3	1 920	3 213	570	30	3,37
SatoFull	H_2O_2	4	1 920	3 213	549	80	3,50
SatoFull	H_2O_2	5	1 920	3 213	543	37	3,54
SatoCore	Cd	1	1 152	1 052	499	14	2,31
SatoCore	Cd	2	1 152	1 052	504	11	2,29
SatoCore	Cd	3	1 152	1 052	498	19	2,31
SatoCore	Cd	4	1 152	1 052	495	17	2,33
SatoCore	Cd	5	1 152	1 052	496	16	2,32
SatoCore	Cd	6	1 152	1 052	497	16	2,32
SatoCore	Cd	7	1 152	1 052	511	12	2,25
SatoCore	Cd	8	1 152	1 052	497	19	2,32
SatoCore	Cd	9	1 152	1 052	492	15	2,34
SatoCore	H_2O_2	1	1 152	1 052	464	21	2,48
SatoCore	H_2O_2	2	1 152	1 052	466	18	2,47
SatoCore	H_2O_2	3	1 152	1 052	466	12	2,47
SatoCore	H_2O_2	4	1 152	1 052	466	15	2,47
SatoCore	H_2O_2	5	1 152	1 052	465	18	2,48

Tab. I.1: **Décomposition en modules par MCL.**

I.2 Obtention d'une liste de protéines d'intérêt

Les protéines d'intérêt sont impliquées notamment dans l'homéostasie du fer. Elles ont été obtenues sur la base des annotations du génome, de recherches de motifs connus pour être impliqués dans la coordination des centre Fe-S et de la connaissance de la littérature. D'autres protéines ont été ajoutées car elles sont impliquées dans des processus d'intérêt pour le laboratoire, par exemple la division cellulaire.

Les protéines d'intérêt sont identifiées par l'identifiant dans la base de données **Uniprot**. Le gène correspondant est indiqué par l'identifiant dans la base de données **Cyanobase**. Quand le gène possède un nom d'usage, il est indiqué dans la colonne **Gène**. La description de la protéine dans la base de données Uniprot est ajoutée dans la colonne **Description**.

Cyanobase	Uniprot	Gène	Description
sll0031	Q55456		Sll0031 protein
sll0037	Q55451	cbiX	Sirohydrochlorin cobaltochelatase
sll0067	Q55139	gst	Glutathione S-transferase
sll0088	Q55789		Sll0088 protein
sll0098	Q55880		UPF0063 protein sll0098

Suite sur la page suivante

Tab. I.2 – suite de la page précédente

Cyanobase	Uniprot	Gène	Description
sll0169	Q55559		Sll0169 protein
sll0170	P22358	dnaK2	Chaperone protein dnaK2
sll0217	P72723	dfa2	diflavin flavoprotein A 2
sll0219	P72721	dfa4	diflavin flavoprotein A 4
sll0254	P73872		Sll0254 protein
sll0258	Q55013	psbV	Cytochrome c-550 precursor
sll0264	P74403		Sll0264 protein
sll0520	P26525	ndhI	NAD(P)H-quinone oxidoreductase subunit I
sll0550	Q55393	dfa1	Diflavin flavoprotein A 1
sll0554	Q55389	ftrC	Ferredoxin-thioredoxin reductase, catalytic chain
sll0567	P74739	fur	Ferric uptake regulation protein
sll0594	Q55854	cysR	Regulatory protein cysR homolog
sll0662	Q55980		Ferredoxin
sll0704	P72676	nifS	NifS protein
sll0741	P52965	nifJ	pyruvate-flavodoxin oxidoreductase
sll0757	Q55621	purF	Amidophosphoribosyltransferase precursor
sll0823	Q55431	sdhB	Succinate dehydrogenase iron-sulfur protein
sll0849	P09192	psbD	Photosystem II D2 protein
sll0868	P73572	lipA2	Lipoyl synthase 2
sll0897	Q55505	dnaJ1	Chaperone protein dnaJ 1
sll0996	P73127		UPF0004 protein sll0996
sll1019	P72933	gloB	hydroxyacylglutathione hydrolase
sll1079	P73268	hypB	Hydrogenase expression/formation protein ; HypB
sll1147	P73795		Uncharacterized protein sll1147
sll1161	P73774		Adenylate cyclase
sll1182	P74174	petC	Cytochrome b6/f complex iron-sulfur subunit
sll1184	P72849	pbsA1	Heme oxygenase 1
sll1205	P72595	pchR	Regulatory protein ; PchR
sll1220	P74025		Potential NAD-reducing hydrogenase subunit
sll1223	P74022	hoxU	Hydrogenase subunit
sll1242	P42349		methyltransferase sll1242
sll1245	P42351	cytM	Cytochrome c-553-like precursor
sll1285	P73191	cofG	FO synthase subunit 1
sll1297	P73170	pobA	Phenoxybenzoate dioxygenase
sll1316	P26290	petC2	Cytochrome b6-f complex iron-sulfur subunit 2
sll1317	P26287	petA	Apocytochrome f precursor
sll1341	P24602	bfr	Bacterioferritin
sll1342	P80505	gap2	Glyceraldehyde-3-phosphate dehydrogenase 2
sll1348	P73526		Sll1348 protein
sll1382	P74159	petF	Ferredoxin
sll1408	P72600	pcrR	Regulatory protein ; PcrR
sll1432	P74218	hypB	hydrogenase nickel incorporation protein hypB
sll1454	P73448	narB	Nitrate reductase
sll1470	P54384	leuC	3-isopropylmalate dehydratase large subunit
sll1499	P55038	gltS	Ferredoxin-dependent glutamate synthase 2

Suite sur la page suivante

Tab. I.2 – suite de la page précédente

Cyanobase	Uniprot	Gène	Description
sll1502	P55037	gltB	Ferredoxin-dependent glutamate synthase 1
sll1521	P74373	dfa3	diflavin flavoprotein A 3
sll1545	P74665	gst1	Glutathione S-transferase
sll1584	P73195		Sll1584 protein
sll1625	P73723	sdhB	Succinate dehydrogenase iron-sulphur protein subunit
sll1633	P73456	ftsZ	Cell division protein ftsZ
sll1659	P72811	cofH	FO synthase subunit 2
sll1766	P73639		protein sll1766
sll1787	P77965	rpoB	DNA-directed RNA polymerase beta chain
sll1796	P46445	petJ	Cytochrome c6 precursor
sll1831	P73119	glcF	Glycolate oxidase subunit (Fe-S) protein
sll1849	P74496		Sll1849 protein
sll1867	P16033		
sll1875	P74133	pbsA2	Heme oxygenase 2
sll1876	P74132	hemN	Oxygen-independent coproporphyrinogen III oxidase
sll1917	P73245		Oxygen-independent coproporphyrinogen III oxidase-like protein sll1917
sll1937	P73084	fur	Ferric uptake regulation protein
sll8031	P17062	ndhK2	NAD(P)H-quinone oxidoreductase subunit K homolog 2
slr0077	Q55793	csd	cysteine desulfurase
slr0082	Q55803		UPF0004 protein slr0082
slr0148	P74447		Ferredoxin
slr0150	P74449	petF	Ferredoxin
slr0236	P72690		Slr0236 protein
slr0309	Q55914		methyltransferase slr0309
slr0342	Q57038	petB	Cytochrome b6
slr0387	Q55602	nifS	NifS protein
slr0452	P74689	ilvD	Dihydroxy-acid dehydratase
slr0574	Q59990	cyp120	cytochrome P450 120
slr0623	P52231	trxA	Thioredoxin
slr0665	P74582	acnB	Aconitate hydratase 2
slr0749	P28373	chlL	Light-independent protochlorophyllide reductase iron-sulfur ATP-binding protein
slr0839	P54225	hemH	Ferrochelatase
slr0846	Q55433		HTH-type transcriptional regulator slr0846
slr0884	P49433	gap1	Glyceraldehyde-3-phosphate dehydrogenase 1
slr0901	Q55369	moaA	Molybdenum cofactor biosynthesis protein A
slr0905	Q55373	bchE	Magnesium-protoporphyrin IX monomethyl ester [oxidative] cyclase
slr0927			
slr0963	P72854	sir	Sulfite reductase [ferredoxin]
slr1137	Q06473	ctaD	Cytochrome c oxidase subunit 1

Suite sur la page suivante

Tab. I.2 – suite de la page précédente

Cyanobase	Uniprot	Gène	Description
slr1167	P74246		oxidoreductase slr1167
slr1181	P07826	psbA1	Photosystem Q(B) protein
slr1185	P74714	petC1	Cytochrome b6-f complex iron-sulfur subunit 1
slr1233	P73479	frdA	Succinate dehydrogenase flavoprotein subunit
slr1259	P73799		Slr1259 protein
slr1265	P74177	rpoC1	DNA-directed RNA polymerase gamma chain
slr1280	P19050	ndhK	NAD(P)H-quinone oxidoreductase subunit K
slr1311	P16033	psbA2	Photosystem Q(B) protein
slr1364	P73538	bioB	Biotin synthase
slr1464	P74557		UPF0026 protein slr1464
slr1489	P72608	pchR	Regulatory protein ; PchR
slr1516	P77968	sodB	Superoxide dismutase [Fe]
slr1520	P73964		Slr1520 protein
slr1529	P74210		Nitrogen assimilation regulatory protein
slr1549	P73441	def	Peptide deformylase
slr1562	P74593		glutaredoxin slr1562
slr1598	P72980	lipA1	Lipoyl synthase 1
slr1738	P73729		Slr1738 protein
slr1822	P73715	nth	Endonuclease III
slr1828	P73388	petF	Ferredoxin
slr1834	P29254	psaA	Photosystem I P700 chlorophyll a apoprotein A1
slr1835	P29255	psaB	Photosystem I P700 chlorophyll a apoprotein A2
slr1846	P73056		monothiol glutaredoxin ycf64-like
slr1849	P73059	merA	Mercuric reductase
slr2033	P73068	rub	Rubredoxin
slr2059	P73811		Ferredoxin
slr2073	P73376		Ycf50-like protein
slr2097	P73925	glbN	Cyanoglobin
smr0006	P09191	psbF	Cytochrome b559 subunit beta
ssl0020	P27320	petF	Ferredoxin-1
ssl0563	P32422	psaC	Photosystem I iron-sulfur center
ssl2250	P73484		Glycoprotein 64
ssl2502			
ssl2559	P73171		Ferredoxin
ssl2667	P74558		NifU protein
ssl3044	P74283		Hydrogenase component
ssr1041	P74801		Ssr1041 protein
ssr2061	P73492		glutaredoxin ssr2061
ssr3184	P73649		Ferredoxin
ssr3451	P09190	psbE	Cytochrome b559 subunit alpha

Tab. I.2: **Liste des protéines d'intérêt.**

I.3 Analyse des transitions

Pour chaque transition d'un temps caractérisé par **T1** modules, au suivant, caractérisé par **T2** modules, le nombre d'événements identifiés est indiqué dans la colonne **Events**. Nous indiquons ensuite les nombre d'événements identifiés dans chacune des catégories suivantes : conservation **P**, regroupement **M**, division **S**, apparition **A**, disparition **D**. La notation Tn>i->Tn+1>i indique une transition du temps Tn au temps suivant Tn+1, seuls les modules de taille supérieure à i étant considérés.

Transition	**T1**	**T2**	**Events**	**P**	**M**	**S**	**A**	**D**
InteroFull - Cd								
T1>0->T2>0	306	304	347	151	34	22	65	75
T2>0->T3>0	304	267	332	153	28	15	80	56
T3>0->T4>0	267	297	337	155	20	16	56	90
T4>0->T5>0	297	302	333	174	25	21	52	61
T5>0->T6>0	302	310	346	176	30	15	51	74
T6>0->T7>0	310	304	331	179	35	17	44	56
T7>0->T8>0	304	301	334	179	27	19	52	57
T8>0->T9>0	301	297	339	153	35	18	60	73
T1>1->T2>1	179	167	199	51	29	19	51	49
T2>1->T3>1	167	140	188	43	26	12	60	47
T3>1->T4>1	140	179	202	55	13	18	41	75
T4>1->T5>1	179	168	191	70	26	17	40	38
T5>1->T6>1	168	181	200	71	26	13	32	58
T6>1->T7>1	181	176	196	75	28	15	35	43
T7>1->T8>1	176	175	196	69	25	18	39	45
T8>1->T9>1	175	160	194	51	31	14	48	50
T1>2->T2>2	116	115	142	33	21	7	34	47
T2>2->T3>2	115	95	127	35	17	7	39	29
T3>2->T4>2	95	122	132	29	12	16	26	49
T4>2->T5>2	122	116	134	36	18	16	34	30
T5>2->T6>2	116	112	125	43	24	6	19	33
T6>2->T7>2	112	130	125	45	15	21	16	28
T7>2->T8>2	130	116	127	43	24	14	25	21
T8>2->T9>2	116	106	131	27	24	8	33	39
InteroFull - H_2O_2								
T1>0->T2>0	265	277	319	137	22	21	63	76
T2>0->T3>0	277	285	359	137	21	12	86	103
T3>0->T4>0	285	275	341	133	30	13	79	86
T4>0->T5>0	275	288	326	129	29	25	63	80
T1>1->T2>1	166	166	204	58	19	16	54	57
T2>1->T3>1	166	182	241	59	15	9	68	90
T3>1->T4>1	182	169	223	48	30	10	64	71
T4>1->T5>1	169	172	205	54	24	17	50	60
T1>2->T2>2	110	105	134	33	16	8	37	40
T2>2->T3>2	105	115	153	31	12	6	44	60

Suite sur la page suivante

Tab. I.3 – suite de la page précédente

Transition	T1	T2	Events	P	M	S	A	D
T3>2->T4>2	115	110	137	30	21	8	35	43
T4>2->T5>2	110	111	136	35	13	12	37	39
SatoFull - Cd								
T1>0->T2>0	623	613	637	447	49	27	51	63
T2>0->T3>0	613	616	629	452	44	30	43	60
T3>0->T4>0	616	625	632	477	37	29	36	53
T4>0->T5>0	625	623	625	479	43	29	31	43
T5>0->T6>0	623	623	632	500	33	24	33	42
T6>0->T7>0	623	624	618	503	38	25	19	33
T7>0->T8>0	624	617	620	487	40	27	30	36
T8>0->T9>0	617	609	607	465	47	30	28	37
T1>1->T2>1	326	315	339	150	49	27	51	62
T2>1->T3>1	315	319	330	154	44	31	42	59
T3>1->T4>1	319	327	334	180	37	29	36	52
T4>1->T5>1	327	325	327	183	43	28	30	43
T5>1->T6>1	325	326	335	204	32	24	33	42
T6>1->T7>1	326	326	320	204	39	25	19	33
T7>1->T8>1	326	319	322	187	41	27	30	37
T8>1->T9>1	319	312	309	166	47	31	28	37
T1>2->T2>2	235	236	246	103	40	21	31	51
T2>2->T3>2	236	236	245	111	33	25	34	42
T3>2->T4>2	236	246	244	120	31	28	26	39
T4>2->T5>2	246	232	238	140	35	15	21	27
T5>2->T6>2	232	247	248	149	17	24	25	33
T6>2->T7>2	247	238	236	133	38	20	18	27
T7>2->T8>2	238	233	237	134	31	19	23	30
T8>2->T9>2	233	239	231	133	26	28	20	24
SatoFull - H_2O_2								
T1>0->T2>0	557	561	576	386	48	30	45	67
T2>0->T3>0	561	570	577	380	53	34	41	69
T3>0->T4>0	570	549	572	385	49	32	55	51
T4>0->T5>0	549	543	572	374	47	26	55	70
T1>1->T2>1	335	339	354	166	48	29	44	67
T2>1->T3>1	339	348	355	160	53	33	40	69
T3>1->T4>1	348	328	351	165	48	32	55	51
T4>1->T5>1	328	321	350	153	47	26	55	69
T1>2->T2>2	240	248	253	119	34	24	29	47
T2>2->T3>2	248	244	243	103	43	30	29	38
T3>2->T4>2	244	236	241	111	36	28	33	33
T4>2->T5>2	236	234	240	106	43	19	25	47
SatoCore - Cd								
T1>0->T2>0	499	504	493	424	22	21	10	16
T2>0->T3>0	504	498	492	422	26	18	12	14
T3>0->T4>0	498	495	485	428	25	15	5	12

Suite sur la page suivante

Tab. I.3 – suite de la page précédente

Transition	T1	T2	Events	P	M	S	A	D
T4>0->T5>0	495	496	485	432	21	16	5	11
T5>0->T6>0	496	497	487	446	15	15	5	6
T6>0->T7>0	497	511	495	449	11	21	5	9
T7>0->T8>0	511	497	497	451	21	9	9	7
T8>0->T9>0	497	492	486	441	18	13	7	7
T1>1->T2>1	288	292	282	214	22	20	10	16
T2>1->T3>1	292	286	280	210	26	18	12	14
T3>1->T4>1	286	283	274	217	25	14	5	13
T4>1->T5>1	283	285	274	222	20	16	5	11
T5>1->T6>1	285	286	276	235	15	15	5	6
T6>1->T7>1	286	298	284	240	11	19	5	9
T7>1->T8>1	298	283	283	236	22	9	9	7
T8>1->T9>1	283	279	273	229	17	13	7	7
T1>2->T2>2	149	151	149	109	9	12	10	9
T2>2->T3>2	151	148	144	113	12	9	5	5
T3>2->T4>2	148	141	141	112	10	8	8	3
T4>2->T5>2	141	155	146	116	6	11	2	11
T5>2->T6>2	155	150	148	121	10	8	6	3
T6>2->T7>2	150	149	145	120	7	10	6	2
T7>2->T8>2	149	149	146	116	11	7	4	8
T8>2->T9>2	149	145	145	113	11	7	7	7
SatoCore - H_2O_2								
T1>0->T2>0	464	466	450	374	27	26	10	13
T2>0->T3>0	466	466	457	383	26	20	11	17
T3>0->T4>0	466	466	457	387	23	21	12	14
T4>0->T5>0	466	465	458	391	23	18	11	15
T1>1->T2>1	301	305	290	218	24	25	10	13
T2>1->T3>1	305	303	294	224	26	19	10	15
T3>1->T4>1	303	305	294	228	21	22	11	12
T4>1->T5>1	305	305	297	229	23	19	11	15
T1>2->T2>2	156	163	153	110	12	16	6	9
T2>2->T3>2	163	162	156	117	13	13	7	6
T3>2->T4>2	162	160	153	123	14	9	2	5
T4>2->T5>2	160	160	152	114	14	13	5	6

Tab. I.3: **Évolution des modules d'un temps à l'autre.**

Quelques liens

La liste suivante indique des liens qui ont été largement utilisés au cours de ce travail de thèse et peuvent être utiles au lecteur.

- Cyanobase : http ://bacteria.kazusa.or.jp/cyanobase/index.html
- DIP : http ://dip.doe-mbi.ucla.edu/
- HUPO : http ://www.hupo.org/
- IntAct : http ://www.ebi.ac.uk/intact/
- Integr8 : http ://www.ebi.ac.uk/integr8/
- KEGG : http ://www.genome.jp/kegg/
- MINT : http ://mint.bio.uniroma2.it/mint/
- OBO : http ://www.obofoundry.org/
- OLS : http ://www.ebi.ac.uk/ontology-lookup/browse.do ?ontName=MI
- PDB : http ://www.rcsb.org/pdb/home/home.do
- Pfam : http ://www.sanger.ac.uk/Software/Pfam/
- PSI : http ://www.psidev.info/
- The Bioconductor Project : http ://www.bioconductor.org

Références

[Albert *et al.*, 2000] R. Albert, H. Jeong et A. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794) :378–82, juillet 2000.

[Aloy, 2007] P. Aloy. Shaping the future of interactome networks. *Genome Biol*, 8(10) :316, novembre 2007.

[Aloy et Russell, 2004] P. Aloy et R. B. Russell. Ten thousand interactions for the molecular biologist. *Nat Biotechnol*, 22(10) :1317–21, octobre 2004.

[Altschul *et al.*, 1990] S. Altschul, W. Gish, W. Miller, E. Myers et D. Lipman. Basic local alignment search tool. *J. Mol. Biol*, 215 :403–410, janvier 1990.

[Arifuzzaman *et al.*, 2006] M. Arifuzzaman, M. Maeda, A. Itoh, K. Nishikata, C. Takita, R. Saito, T. Ara, K. Nakahigashi, H.-C. Huang, A. Hirai, K. Tsuzuki, S. Nakamura, M. Altaf-Ul-Amin, T. Oshima, T. Baba, N. Yamamoto, T. Kawamura, T. Ioka-Nakamichi, M. Kitagawa, M. Tomita, S. Kanaya, C. Wada et H. Mori. Large-scale identification of protein-protein interaction of escherichia coli k-12. *Genome Res*, 16(5) :686–91, mai 2006.

[Ashburner *et al.*, 2000] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin et G. Sherlock. Gene ontology : tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1) :25–9, mai 2000.

[Assenov *et al.*, 2007] Y. Assenov, F. Ramírez, S. Schelhorn, T. Lengauer et M. Albrecht. Computing topological parameters of biological networks. *Bioinformatics*, 24(2) :282–284, novembre 2007.

[Aude, 1999] J. C. Aude. Analyse de génomes microbiens : apports de la classification pyramidale. *Thèse de l'Université Paris IX*, 1999.

[Bachmann, 2003] T. Bachmann. Transforming cyanobacteria into bioreporters of biological relevance. *Trends Biotechnol*, 21(6) :247–9, juin 2003.

[Bader *et al.*, 2003] G. D. Bader, D. Betel et C. W. V. Hogue. Bind : the biomolecular interaction network database. *Nucleic Acids Res*, 31(1) :248–50, janvier 2003.

[Bader *et al.*, 2001] G. D. Bader, I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson et C. W. Hogue. Bind–the biomolecular interaction network database. *Nucleic Acids Res*, 29(1) :242–5, janvier 2001.

[Bader et Hogue, 2000] G. D. Bader et C. W. Hogue. Bind–a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics*, 16(5) :465–77, mai 2000.

[Bader et Hogue, 2003] G. D. Bader et C. W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4 :2, janvier 2003.

[Bandyopadhyay *et al.*, 2006] S. Bandyopadhyay, R. Sharan et T. Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Res*, 16(3) :428–35, mars 2006.

[Bar-Joseph *et al.*, 2003] Z. Bar-Joseph, G. K. Gerber, D. K. Gifford, T. S. Jaakkola et I. Simon. Continuous representations of time-series gene expression data. *J Comput Biol*, 10(3-4) :341–56, janvier 2003.

[Barabasi et Albert, 1999] A. Barabasi et R. Albert. Emergence of scaling in random networks. *Science*, 286(5439) :509–12, octobre 1999.

[Batada *et al.*, 2006] N. N. Batada, T. Reguly, A. Breitkreutz, L. Boucher, B.-J. Breitkreutz, L. D. Hurst et M. Tyers. Stratus not altocumulus : a new view of the yeast protein interaction network. *PLoS Biol*, 4(10) :e317, octobre 2006.

[Baudot *et al.*, 2008] A. Baudot, J. Angelelli, A. Guenoche, B. Jacq et C. Brun. Defining a modular signalling network from the fly interactome. *BMC Systems Biology*, 2(1) :45, mai 2008.

[Baudouin-Cornu *et al.*, 2001] P. Baudouin-Cornu, Y. Surdin-Kerjan, P. Marlière et D. Thomas. Molecular evolution of protein atomic composition. *Science*, 293(5528) :297–300, juillet 2001.

[Ben-Hur et Noble, 2005] A. Ben-Hur et W. S. Noble. Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21 Suppl 1 :i38–46, juin 2005.

[Benjamini et Yekutieli, 2001] Y. Benjamini et D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4) :1165–1188, 2001.

[Bennetzen et Hall, 1982] J. L. Bennetzen et B. D. Hall. Codon selection in yeast. *J Biol Chem*, 257(6) :3026–31, mars 1982.

[Bertin *et al.*, 2007] N. Bertin, N. Simonis, D. Dupuy, M. E. Cusick, J.-D. J. Han, H. B. Fraser, F. P. Roth et M. Vidal. Confirmation of organized modularity in the yeast interactome. *PLoS Biol*, 5(6) :e153, juin 2007.

[Bertrand et Diday, 1990] P. Bertrand et E. Diday. Une généralisation des arbres hiérarchiques : les représentations pyramidales. *Rev. Statistique Appliquée*, 38(3) :53–78, septembre 1990.

[Blatt *et al.*, 1996] M. Blatt, S. Wiseman et E. Domany. Superparamagnetic clustering of data. *Phys Rev Lett*, 76(18) :3251–3254, avril 1996.

[Bolshakova et Azuaje, 2003] N. N. Bolshakova et F. F. Azuaje. Cluster validation techniques for genome expression data. *Elsevier Science*, 83 :825–833, avril 2003.

[Boxem *et al.*, 2008] M. Boxem, Z. Maliga, N. Klitgord, N. Li, I. Lemmens, M. Mana, L. de Lichtervelde, J. D. Mul, D. van de Peut, M. Devos, N. Simonis, M. A. Yildirim, M. Cokol, H.-L. Kao, A.-S. de Smet, H. Wang, A.-L. Schlaitz, T. Hao, S. Milstein, C. Fan, M. Tipsword, K. Drew, M. Galli, K. Rhrissorrakrai, D. Drechsel, D. Koller, F. P. Roth, L. M. Iakoucheva, A. K. Dunker, R. Bonneau, K. C. Gunsalus, D. E. Hill, F. Piano, J. Tavernier, S. van den Heuvel, A. A. Hyman et M. Vidal. A protein domain-based interactome network for c. elegans early embryogenesis. *Cell*, 134(3) :534–45, août 2008.

[Bradford et Westhead, 2005] J. R. Bradford et D. R. Westhead. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8) :1487–94, avril 2005.

[Breitkreutz *et al.*, 2007] B. Breitkreutz, C. Stark, T. Reguly, L. Boucher, A. Breitkreutz, M. Livstone, R. Oughtred, D. Lackner, J. Bähler, V. Wood, K. Dolinski et M. Tyers. The biogrid interaction database : 2008 update. *Nucleic Acids Res*, novembre 2007.

[Brohée et van Helden, 2006] S. Brohée et J. van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7 :488, janvier 2006.

[Brown et Jurisica, 2005] K. R. Brown et I. Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9) :2076–82, mai 2005.

[Brown et Jurisica, 2007] K. R. Brown et I. Jurisica. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol*, 8(5) :R95, janvier 2007.

[Brown *et al.*, 2000] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares et D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA*, 97(1) :262–7, janvier 2000.

[Brun *et al.*, 2003] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guénoche et B. Jacq. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol*, 5(1) :R6, janvier 2003.

[Burger et van Nimwegen, 2008] L. Burger et E. van Nimwegen. Accurate prediction of protein-protein interactions from sequence alignments using a bayesian method. *Mol Syst Biol*, 4 :165, janvier 2008.

[Calvano *et al.*, 2005] S. E. Calvano, W. Xiao, D. R. Richards, R. M. Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, S. K. Tschoeke, C. Miller-Graziano, L. L. Moldawer, M. N. Mindrinos, R. W. Davis, R. G. Tompkins, S. F. Lowry, Inflamm et H. R. to Injury Large Scale Collab. Res. Program. A network-based analysis of systemic inflammation in humans. *Nature*, 437(7061) :1032–7, octobre 2005.

[Cech et Rubin, 2004] T. R. Cech et G. M. Rubin. Nurturing interdisciplinary research. *Nat Struct Mol Biol*, 11(12) :1166–9, décembre 2004.

[Chen et Liu, 2005] X.-W. Chen et M. Liu. Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, 21(24) :4394–400, décembre 2005.

[Chiang *et al.*, 2007] T. Chiang, D. Scholtens, D. Sarkar, R. Gentleman et W. Huber. Coverage and error models of protein-protein interaction data by directed graph analysis. *Genome Biol*, 8(9) :R186, septembre 2007.

[Clauset *et al.*, 2008] A. Clauset, C. Moore et M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191) :98–101, mai 2008.

[Claverie, 1999] J. M. Claverie. Computational methods for the identification of differential and coordinated gene expression. *Hum Mol Genet*, 8(10) :1821–32, janvier 1999.

[Cleveland et Devlin, 1988] W. S. Cleveland et S. J. Devlin. Locally-weighted regression : An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83 :596–610, 1988.

[Cokus *et al.*, 2007] S. Cokus, S. Mizutani et M. Pellegrini. An improved method for identifying functionally linked proteins using phylogenetic profiles. *BMC Bioinformatics*, 8 Suppl 4 :S7, janvier 2007.

[Dandekar *et al.*, 1998] T. Dandekar, B. Snel, M. Huynen et P. Bork. Conservation of gene order : a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23(9) :324–8, septembre 1998.

[Deng *et al.*, 2002] M. Deng, S. Mehta, F. Sun et T. Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, 12(10) :1540–8, octobre 2002.

[DeRisi *et al.*, 1997] J. L. DeRisi, V. R. Iyer et P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338) :680–6, octobre 1997.

[Descorps-Declère *et al.*, 2008] S. Descorps-Declère, F. Lemoine, Q. Sculo, O. Lespinet et B. Labedan. The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species. *Biochimie*, 90(4) :595–608, avril 2008.

[Domain *et al.*, 2004] F. Domain, L. Houot, F. Chauvat et C. Cassier-Chauvat. Function and regulation of the cyanobacterial genes lexa, reca and ruvb : Lexa is critical to the survival of cells facing inorganic carbon starvation. *Mol Microbiol*, 53(1) :65–80, juillet 2004.

[Doolittle, 1981] R. F. Doolittle. Similar amino acid sequences : chance or common ancestry ? *Science*, 214(4517) :149–59, octobre 1981.

[Eisen *et al.*, 1998] M. B. Eisen, P. T. Spellman, P. O. Brown et D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95(25) :14863–8, décembre 1998.

[Enright *et al.*, 2002] A. J. Enright, S. V. Dongen et C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7) :1575–84, avril 2002.

[Enright *et al.*, 1999] A. J. Enright, I. Iliopoulos, N. C. Kyrpides et C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757) :86–90, novembre 1999.

[Fariselli *et al.*, 2002] P. Fariselli, F. Pazos, A. Valencia et R. Casadio. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem*, 269(5) :1356–61, mars 2002.

[Fauchon *et al.*, 2002] M. Fauchon, G. Lagniel, J. C. Aude, L. Lombardia, P. Soularue, C. Petat, G. Marguerie, A. Sentenac, M. Werner et J. Labarre. Sulfur sparing in the yeast proteome in response to sulfur demand. *Mol Cell*, 9(4) :713–23, avril 2002.

[Fields et Song, 1989] S. Fields et O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230) :245–6, juillet 1989.

[Finn *et al.*, 2005] R. D. Finn, M. Marshall et A. Bateman. ipfam : visualization of protein-protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics*, 21(3) :410–2, février 2005.

[Fitch, 1970] W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19(2) :99–113, juin 1970.

[Formstecher *et al.*, 2005] E. Formstecher, S. Aresta, V. Collura, A. Hamburger, A. Meil, A. Trehin, C. Reverdy, V. Betin, S. Maire, C. Brun, B. Jacq, M. Arpin, Y. Bellaiche, S. Bellusci, P. Benaroch, M. Bornens, R. Chanet, P. Chavrier, O. Delattre, V. Doye, R. Fehon, G. Faye, T. Galli, J.-A. Girault, B. Goud, J. de Gunzburg, L. Johannes, M.-P. Junier, V. Mirouse, A. Mukherjee, D. Papadopoulo, F. Perez, A. Plessis, C. Rossé, S. Saule, D. Stoppa-Lyonnet, A. Vincent, M. White, P. Legrain, J. Wojcik, J. Camonis et L. Daviet. Protein interaction mapping : a drosophila case study. *Genome Res*, 15(3) :376–84, mars 2005.

[Fraser *et al.*, 2002] H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe et M. W. Feldman. Evolutionary rate in the protein interaction network. *Science*, 296(5568) :750–2, avril 2002.

[Fraser *et al.*, 2004] H. B. Fraser, A. E. Hirsh, D. P. Wall et M. B. Eisen. Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci USA*, 101(24) :9033–8, juin 2004.

[Fromont-Racine *et al.*, 2002] M. Fromont-Racine, J.-C. Rain et P. Legrain. Building protein-protein networks by two-hybrid mating strategy. *Meth Enzymol*, 350 :513–24, janvier 2002.

[Gandhi *et al.*, 2006] T. K. B. Gandhi, J. Zhong, S. Mathivanan, L. Karthick, K. N. Chandrika, S. S. Mohan, S. Sharma, S. Pinkert, S. Nagaraju, B. Periaswamy, G. Mishra, K. Nandakumar, B. Shen, N. Deshpande, R. Nayak, M. Sarker, J. D. Boeke, G. Parmigiani, J. Schultz, J. S. Bader et A. Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3) :285–93, mars 2006.

[Garrity et Lilburn, 2005] G. M. Garrity et T. G. Lilburn. Self-organizing and self-correcting classifications of biological data. *Bioinformatics*, 21(10) :2309–14, mai 2005.

[Gavin *et al.*, 2002] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer et G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868) :141–7, janvier 2002.

[Ge *et al.*, 2001] H. Ge, Z. Liu, G. M. Church et M. Vidal. Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nat Genet*, 29(4) :482–6, décembre 2001.

[Gentleman *et al.*, 2005] R. Gentleman, V. Carey, W. Huber, R. Irizarry et S. Dudoit. Bioinformatics and computational biology solutions using r and bioconductor. *Springer Series in Statistics for Biology and Health*, 2005.

[Gentleman et Huber, 2007] R. Gentleman et W. Huber. Making the most of high-throughput protein-interaction data. *Genome Biol*, 8(10) :112, janvier 2007.

[Gentleman *et al.*, 2004] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang et J. Zhang. Bioconductor : open software development for computational biology and bioinformatics. *Genome Biol*, 5(10) :R80, janvier 2004.

[Gerstein *et al.*, 2002] M. Gerstein, N. Lan et R. Jansen. Proteomics. integrating interactomes. *Science*, 295(5553) :284–7, janvier 2002.

[Gertz *et al.*, 2003] J. Gertz, G. Elfond, A. Shustrova, M. Weisinger, M. Pellegrini, S. Cokus et B. Rothschild. Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*, 19(16) :2039–45, novembre 2003.

[Geurts *et al.*, 2007] P. Geurts, N. Touleimat, M. Dutreix et F. d'Alché Buc. Inferring biological networks with output kernel trees. *BMC Bioinformatics*, 8 Suppl 2 :S4, janvier 2007.

[Ghavi-Helm *et al.*, 2008] Y. Ghavi-Helm, M. Michaut, J. Acker, J.-C. Aude, P. Thuriaux, M. Werner et J. Soutourina. Genome-wide location analysis reveals a role of tfiis in rna polymerase iii transcription. *Genes Dev*, 22(14) :1934–47, juillet 2008.

[Giot *et al.*, 2003] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant et J. M. Rothberg. A protein interaction map of drosophila melanogaster. *Science*, 302(5651) :1727–36, décembre 2003.

[Goh *et al.*, 2000] C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther et F. E. Cohen. Co-evolution of proteins with their interaction partners. *J Mol Biol*, 299(2) :283–93, juin 2000.

[Goldberg et Roth, 2003] D. S. Goldberg et F. P. Roth. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci USA*, 100(8) :4372–6, avril 2003.

[Goll et Uetz, 2006] J. Goll et P. Uetz. The elusive yeast interactome. *Genome Biol*, 7(6) :223, janvier 2006.

[Gong *et al.*, 2005] R. Gong, Y. Ding, H. Liu, Q. Chen et Z. Liu. Lead biosorption and desorption by intact and pretreated spirulina maxima biomass. *Chemosphere*, 58(1) :125–30, janvier 2005.

[Gordon, 1996] A. D. Gordon. Clustering and classification. *World Scientific*, chapter Hierachical Classification :65–121, 1996.

[Gray, 1993] M. W. Gray. Origin and evolution of organelle genomes. *Curr Opin Genet Dev*, 3(6) :884–90, décembre 1993.

[Greller et Tobin, 1999] L. D. Greller et F. L. Tobin. Detecting selective expression of genes and proteins. *Genome Res*, 9(3) :282–96, mars 1999.

[Grigoriev, 2001] A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale : analysis of the bacteriophage t7 and the yeast saccharomyces cerevisiae. *Nucleic Acids Res*, 29(17) :3513–9, septembre 2001.

[Guruprasad *et al.*, 1990] K. Guruprasad, B. V. Reddy et M. W. Pandit. Correlation between stability of a protein and its dipeptide composition : a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng*, 4(2) :155–61, décembre 1990.

[Hakes *et al.*, 2007] L. Hakes, S. C. Lovell, S. G. Oliver et D. L. Robertson. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc Natl Acad Sci USA*, 104(19) :7999–8004, mai 2007.

[Hakes *et al.*, 2008] L. Hakes, J. W. Pinney, D. L. Robertson et S. C. Lovell. Protein-protein interaction networks and biology-what's the connection ? *Nat Biotechnol*, 26(1) :69–72, janvier 2008.

[Han *et al.*, 2004a] D.-S. Han, H.-S. Kim, W.-H. Jang, S.-D. Lee et J.-K. Suh. Prespi : a domain combination based prediction system for protein-protein interaction. *Nucleic Acids Res*, 32(21) :6312–20, janvier 2004.

[Han *et al.*, 2004b] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth et M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995) :88–93, juillet 2004.

[Han *et al.*, 2005] J.-D. J. Han, D. Dupuy, N. Bertin, M. E. Cusick et M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol*, 23(7) :839–44, juillet 2005.

[Harris *et al.*, 2004] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White et G. O. Consortium. The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, 32(Database issue) :D258–61, janvier 2004.

[Hart *et al.*, 2007] G. T. Hart, I. Lee et E. R. Marcotte. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, 8 :236, janvier 2007.

[Hart *et al.*, 2006] G. T. Hart, A. K. Ramani et E. M. Marcotte. How complete are current yeast and human protein-interaction networks ? *Genome Biol*, 7(11) :120, janvier 2006.

[Hartwell *et al.*, 1999] L. H. Hartwell, J. J. Hopfield, S. Leibler et A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl) :C47–52, décembre 1999.

[Heck, 2008] A. J. R. Heck. Native mass spectrometry : a bridge between interactomics and structural biology. *Nat Methods*, 5(11) :927–33, novembre 2008.

[Hermjakob *et al.*, 2004a] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. G. N. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue et R. Apweiler. The hupo psi's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2) :177–83, février 2004.

[Hermjakob *et al.*, 2004b] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman et R. Apweiler. Intact : an open source molecular interaction database. *Nucleic Acids Res*, 32(Database issue) :D452–5, janvier 2004.

[Hirel *et al.*, 1989] P. H. Hirel, M. J. Schmitter, P. Dessen, G. Fayat et S. Blanquet. Extent of n-terminal methionine excision from escherichia coli proteins is governed by the side-chain length of the penultimate amino acid. *Proc Natl Acad Sci USA*, 86(21) :8247–51, novembre 1989.

[Hirsh et Fraser, 2001] A. E. Hirsh et H. B. Fraser. Protein dispensability and rate of evolution. *Nature*, 411(6841) :1046–9, juin 2001.

[Ho *et al.*, 2002] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson,

S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sørensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys et M. Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868) :180–3, janvier 2002.

[Hochberg et Benjamini, 1990] Y. Hochberg et Y. Benjamini. More powerful procedures for multiple significance testing. *Statistics in medicine*, 9(7) :811–8, juillet 1990.

[Houot *et al.*, 2007] L. Houot, M. Floutier, B. Marteyn, M. Michaut, A. Picciocchi, P. Legrain, J. Aude, C. Cassier-Chauvat et F. Chauvat. Cadmium triggers an integrated reprogramming of the metabolism of synechocystis pcc6803, under the control of the slr1738 regulator. *BMC Genomics*, 8(1) :350, octobre 2007.

[Huang *et al.*, 2007a] H. Huang, B. M. Jedynak et J. S. Bader. Where have all the interactions gone ? estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput Biol*, 3(11) :e214, novembre 2007.

[Huang *et al.*, 2007b] T.-W. Huang, C.-Y. Lin et C.-Y. Kao. Reconstruction of human protein interolog network using evolutionary conserved network. *BMC Bioinformatics*, 8 :152, janvier 2007.

[Huang *et al.*, 2004] T.-W. Huang, A.-C. Tien, W.-S. Huang, Y.-C. G. Lee, C.-L. Peng, H.-H. Tseng, C.-Y. Kao et C.-Y. F. Huang. Point : a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, 20(17) :3273–6, novembre 2004.

[Hunter *et al.*, 2001] L. Hunter, R. C. Taylor, S. M. Leach et R. Simon. Gest : a gene expression search tool based on a novel bayesian similarity metric. *Bioinformatics*, 17 Suppl 1 :S115–22, janvier 2001.

[Hunter, 2008] P. Hunter. The paradox of model organisms. the use of model organisms in research will continue despite their shortcomings. *EMBO Rep*, 9(8) :717–20, août 2008.

[Huynen *et al.*, 2000] M. Huynen, B. Snel, W. Lathe et P. Bork. Predicting protein function by genomic context : quantitative evaluation and qualitative inferences. *Genome Res*, 10(8) :1204–10, août 2000.

[Hwang *et al.*, 2006] W. Hwang, Y.-R. Cho, A. Zhang et M. Ramanathan. A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms for molecular biology : AMB*, 1 :24, janvier 2006.

[Ideker *et al.*, 2002] T. Ideker, O. Ozier, B. Schwikowski et A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1 :S233–40, janvier 2002.

[Ito *et al.*, 2001] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori et Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*, 98(8) :4569–74, avril 2001.

[Ito *et al.*, 2002] T. Ito, K. Ota, H. Kubota, Y. Yamaguchi, T. Chiba, K. Sakuraba et M. Yoshida. Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol Cell Proteomics*, 1(8) :561–6, août 2002.

[Itzhaki *et al.*, 2006] Z. Itzhaki, E. Akiva, Y. Altuvia et H. Margalit. Evolutionary conservation of domain-domain interactions. *Genome Biol*, 7(12) :R125, janvier 2006.

[Iyer *et al.*, 1999] V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. Lee, J. M. Trent, L. M. Staudt, J. Hudson, M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein et P. O. Brown. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283(5398) :83–7, janvier 1999.

[Jansen *et al.*, 2003a] R. Jansen, H. J. Bussemaker et M. Gerstein. Revisiting the codon adaptation index from a whole-genome perspective : analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res*, 31(8) :2242–51, avril 2003.

[Jansen *et al.*, 2002] R. Jansen, D. Greenbaum et M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 12(1) :37–46, janvier 2002.

[Jansen *et al.*, 2003b] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt et M. Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644) :449–53, octobre 2003.

[Jensen *et al.*, 2008] L. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork et C. Mering. String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, octobre 2008.

[Jeong *et al.*, 2001] H. Jeong, S. P. Mason, A. L. Barabási et Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833) :41–2, mai 2001.

[Jeong *et al.*, 2003] H. Jeong, Z. Oltvai, A. Barabási et A. Barabsi. Prediction of protein essentiality based on genomic data. *Logo*, janvier 2003.

[Ji *et al.*, 2003] X. Ji, J. Li-Ling et Z. Sun. Mining gene expression data using a novel approach based on hidden markov models. *FEBS Lett*, 542(1-3) :125–31, mai 2003.

[Jones et Thornton, 1997] S. Jones et J. M. Thornton. Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol*, 272(1) :133–43, septembre 1997.

[Jonsson *et al.*, 2006] P. F. Jonsson, T. Cavanna, D. Zicha et P. A. Bates. Cluster analysis of networks generated through homology : automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics*, 7 :2, janvier 2006.

[Jordan *et al.*, 2002] I. K. Jordan, I. B. Rogozin, Y. I. Wolf et E. V. Koonin. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*, 12(6) :962–8, juin 2002.

[Kaneko *et al.*, 1996] T. Kaneko, S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi,

A. Matsuno, A. Muraki, N. Nakazaki, K. Naruo, S. Okumura, S. Shimpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda et S. Tabata. Sequence analysis of the genome of the unicellular cyanobacterium synechocystis sp. strain pcc6803. ii. sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res*, 3(3) :109–36, juin 1996.

[Kemmeren *et al.*, 2002] P. Kemmeren, N. L. van Berkum, J. Vilo, T. Bijma, R. Donders, A. Brazma et F. C. P. Holstege. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell*, 9(5) :1133–43, mai 2002.

[Kerrien *et al.*, 2007a] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler et H. HERMJAKOB. Intact–open source resource for molecular interaction data. *Nucleic Acids Res*, 35(Database issue) :D561–5, janvier 2007.

[Kerrien *et al.*, 2007b] S. Kerrien, S. Orchard, L. Montecchi-Palazzi, B. Aranda, A. Quinn, N. Vinod, G. Bader, I. Xenarios, J. Wojcik, D. Sherman, M. Tyers, J. Salama, S. Moore, A. Ceol, A. Chatr-Aryamontri, M. Oesterheld, V. Stumpflen, L. Salwinski, J. Nerothin, E. Cerami, M. Cusick, M. Vidal, M. Gilson, J. Armstrong, P. Woollard, C. Hogue, D. Eisenberg, G. Cesareni, R. Apweiler et H. Hermjakob. Broadening the horizon - level 2.5 of the hupo-psi format for molecular interactions. *BMC Biol*, 5(1) :44, octobre 2007.

[Kersey *et al.*, 2005] P. Kersey, L. Bower, L. Morris, A. Horne, R. Petryszak, C. Kanz, A. Kanapin, U. Das, K. Michoud, I. Phan, A. Gattiker, T. Kulikova, N. Faruque, K. Duggan, P. Mclaren, B. Reimholz, L. Duret, S. Penel, I. Reuter et R. Apweiler. Integr8 and genome reviews : integrated views of complete genomes and proteomes. *Nucleic Acids Res*, 33(Database issue) :D297–302, janvier 2005.

[Kim *et al.*, 2002] W. K. Kim, J. Park et J. K. Suh. Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair. *Genome informatics International Conference on Genome Informatics*, 13 :42–50, janvier 2002.

[Kim *et al.*, 2008] W.-Y. Kim, S. Kang, B.-C. Kim, J. Oh, S. Cho, J. Bhak et J.-S. Choi. Synechonet : integrated protein-protein interaction database of a model cyanobacterium synechocystis sp. pcc 6803. *BMC Bioinformatics*, 9 Suppl 1 :S20, janvier 2008.

[King *et al.*, 2004] A. D. King, N. Przulj et I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17) :3013–20, novembre 2004.

[Kini et Evans, 1996] R. M. Kini et H. J. Evans. Prediction of potential protein-protein interaction sites from amino acid sequence. identification of a fibrin polymerization site. *FEBS Lett*, 385(1-2) :81–6, avril 1996.

[Klipp *et al.*, 2005] E. Klipp, R. Herwig, A. Kowald, C. Wierling et H. Lehrach. Systems biology in practice. *Wiley-Vch*, Concepts, Implementation and Application, 2005.

[Koike et Takagi, 2004] A. Koike et T. Takagi. Prediction of protein-protein interaction sites using support vector machines. *Protein Eng Des Sel*, 17(2) :165–73, février 2004.

[Komurov et White, 2007] K. Komurov et M. White. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol Syst Biol*, 3 :110, janvier 2007.

[Kyte et Doolittle, 1982] J. Kyte et R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1) :105–32, mai 1982.

[Lappe, 2003] M. Lappe. Novel algorithms for protein interaction networks. *PhD dissertation*, 2003.

[Launay, 2007] G. Launay. Etude theorique des interactions proteine-proteine. *Thèse de doctorat, Ecole Polytechnique*, 2007.

[Lebart *et al.*, 1995] L. Lebart, M. Piron et A. Morineau. Statistique exploratoire multidimensionnelle. *Dunod*, Introduction, 1995.

[Lee *et al.*, 2006] H. Lee, M. Deng, F. Sun et T. Chen. An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics*, 7 :269, janvier 2006.

[Legrain *et al.*, 2001] P. Legrain, J. Wojcik et J. M. Gauthier. Protein–protein interaction maps : a lead towards cellular functions. *Trends Genet*, 17(6) :346–52, juin 2001.

[Lehner et Fraser, 2004] B. Lehner et A. G. Fraser. A first-draft human protein-interaction map. *Genome Biol*, 5(9) :R63, janvier 2004.

[Levy et Pereira-Leal, 2008] E. D. Levy et J. B. Pereira-Leal. Evolution and dynamics of protein interactions and networks. *Curr Opin Struct Biol*, 18(3) :349–57, juin 2008.

[Li *et al.*, 2006] H. Li, C. L. Wood, Y. Liu, T. V. Getchell, M. L. Getchell et A. J. Stromberg. Identification of gene expression patterns using planned linear contrasts. *BMC Bioinformatics*, 7 :245, janvier 2006.

[Li *et al.*, 2004] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D. J. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J.-F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. V. D. Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill et M. Vidal. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657) :540–3, janvier 2004.

[Lin, 1998] D. Lin. An information-theoretic definition of similarity. *Madison, WI*, Morgan Kaufmann :296–304, 1998.

[Livingstone et Barton, 1993] C. D. Livingstone et G. J. Barton. Protein sequence alignments : a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci*, 9(6) :745–56, décembre 1993.

[Long *et al.*, 2001] A. D. Long, H. J. Mangalam, B. Y. Chan, L. Tolleri, G. W. Hatfield et P. Baldi. Improved statistical inference from dna microarray data using analysis

of variance and a bayesian statistical framework. analysis of global gene expression in escherichia coli k12. *J Biol Chem*, 276(23) :19937–44, juin 2001.

[Lubovac *et al.*, 2006] Z. Lubovac, J. Gamalielsson et B. Olsson. Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins*, 64(4) :948–59, septembre 2006.

[Mann *et al.*, 2001] M. Mann, R. C. Hendrickson et A. Pandey. Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem*, 70 :437–73, janvier 2001.

[Marcotte *et al.*, 1999a] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates et D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428) :751–3, juillet 1999.

[Marcotte *et al.*, 1999b] E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates et D. Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757) :83–6, novembre 1999.

[Martin *et al.*, 2002] W. Martin, T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa et D. Penny. Evolutionary analysis of arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA*, 99(19) :12246–51, septembre 2002.

[Matthews *et al.*, 2001] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent et M. Vidal. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res*, 11(12) :2120–6, décembre 2001.

[Mazouni *et al.*, 2004] K. Mazouni, F. Domain, C. Cassier-Chauvat et F. Chauvat. Molecular analysis of the key cytokinetic components of cyanobacteria : Ftsz, zipn and mincde. *Mol Microbiol*, 52(4) :1145–58, mai 2004.

[Mazouni *et al.*, 2003] K. Mazouni, F. Domain, F. Chauvat et C. Cassier-Chauvat. Expression and regulation of the crucial plant-like ferredoxin of cyanobacteria. *Mol Microbiol*, 49(4) :1019–29, août 2003.

[McDermott *et al.*, 2005] J. McDermott, R. Bumgarner et R. Samudrala. Functional annotation from predicted protein interaction networks. *Bioinformatics*, 21(15) :3217–26, août 2005.

[Medvedovic et Sivaganesan, 2002] M. Medvedovic et S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9) :1194–206, septembre 2002.

[Michaut *et al.*, 2008a] M. Michaut, H. Hermjakob, P. Legrain et J.-C. Aude. Synechocystis protein-protein interaction network. *(en préparation)*, 2008.

[Michaut *et al.*, 2008b] M. Michaut, S. Kerrien, L. Montecchi-Palazzi, C. Cassier-Chauvat, F. Chauvat, J.-C. Aude, P. Legrain et H. Hermjakob. Inference of synechocystis protein interaction network. *Proceedings of Journées Ouvertes Biologie Informatique Mathématiques*, pages 99–104, octobre 2008.

[Michaut *et al.*, 2008c] M. Michaut, S. Kerrien, L. Montecchi-Palazzi, F. Chauvat, C. Cassier-Chauvat, J. Aude, P. Legrain et H. HERMJAKOB. Highlights from the fourth international society for computational biology (iscb) student council symposium. *BMC Bioinformatics*, 9(Supplement 10), 2008.

[Michaut *et al.*, 2008d] M. Michaut, S. Kerrien, L. Montecchi-Palazzi, F. Chauvat, C. Cassier-Chauvat, J. Aude, P. Legrain et H. Hermjakob. Interoporc : Automated inference of highly conserved protein interaction networks. *Bioinformatics*, 24(14) :1625–1631, mai 2008.

[Michaut *et al.*, 2007] M. Michaut, S. Kerrien, L. Montecchi-Palazzi, F. Chauvat, C. Cassier-Chauvat, J.-C. Aude, P. Legrain et H. Hermjakob. Inference and validation of synechocystis protein interaction network using orthology. *Proceedings of Journées Ouvertes Biologie Informatique Mathématiques*, pages 229–234, octobre 2007.

[Milenkovic *et al.*, 2008] T. Milenkovic, J. Lai et N. Przulj. Graphcrunch : a tool for large network analyses. *BMC Bioinformatics*, 9(1) :70, janvier 2008.

[Moerschell *et al.*, 1990] R. P. Moerschell, Y. Hosokawa, S. Tsunasawa et F. Sherman. The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine in vivo. processing of altered iso-1-cytochromes c created by oligonucleotide transformation. *J Biol Chem*, 265(32) :19638–43, novembre 1990.

[Mrázek *et al.*, 2001] J. Mrázek, D. Bhaya, A. R. Grossman et S. Karlin. Highly expressed and alien genes of the synechocystis genome. *Nucleic Acids Res*, 29(7) :1590–601, avril 2001.

[Mushegian *et al.*, 1998] A. R. Mushegian, J. R. Garey, J. Martin et L. X. Liu. Large-scale taxonomic profiling of eukaryotic model organisms : a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res*, 8(6) :590–8, juin 1998.

[Myers et Troyanskaya, 2007] C. L. Myers et O. G. Troyanskaya. Context-sensitive data integration and prediction of biological networks. *Bioinformatics*, 23(17) :2322–30, septembre 2007.

[Najafabadi et Salavati, 2008] H. S. Najafabadi et R. Salavati. Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biol*, 9(5) :R87, mai 2008.

[Nakamura *et al.*, 1998] Y. Nakamura, T. Kaneko, M. Hirosawa, N. Miyajima et S. Tabata. Cyanobase, a www database containing the complete nucleotide sequence of the genome of synechocystis sp. strain pcc6803. *Nucleic Acids Research*, 26(1) :63–7, janvier 1998.

[Nakamura *et al.*, 2000] Y. Nakamura, T. Kaneko et S. Tabata. Cyanobase, the genome database for synechocystis sp. strain pcc6803 : status for the year 2000. *Nucleic Acids Research*, 28(1) :72, janvier 2000.

[Neuhäuser et Senske, 2004] M. Neuhäuser et R. Senske. The baumgartner-weiss-schindler test for the detection of differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 20(18) :3553–64, décembre 2004.

[Newton *et al.*, 2001] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner et K. W. Tsui. On differential variability of expression ratios : improving statistical inference about gene expression changes from microarray data. *J Comput Biol*, 8(1) :37–52, janvier 2001.

[Ng *et al.*, 2003a] S.-K. Ng, Z. Zhang et S.-H. Tan. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8) :923–9, mai 2003.

[Ng *et al.*, 2003b] S.-K. Ng, Z. Zhang, S.-H. Tan et K. Lin. Interdom : a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res*, 31(1) :251–4, janvier 2003.

[Nooren et Thornton, 2003] I. M. A. Nooren et J. M. Thornton. Diversity of protein-protein interactions. *EMBO J*, 22(14) :3486–92, juillet 2003.

[Overbeek *et al.*, 1999] R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch et N. Maltsev. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA*, 96(6) :2896–901, mars 1999.

[Owen, 1848] R. Owen. On the archetype and homologies of the vertebrate skeleton. *London*, 1848.

[Pace *et al.*, 1995] C. N. Pace, F. Vajdos, L. Fee, G. Grimsley et T. Gray. How to measure and predict the molar absorption coefficient of a protein. *Protein Sci*, 4(11) :2411–23, novembre 1995.

[Palla *et al.*, 2005] G. Palla, I. Derényi, I. Farkas et T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043) :814–8, juin 2005.

[Pan, 2002] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4) :546–54, avril 2002.

[Pan *et al.*, 2002a] W. Pan, J. Lin et C. T. Le. How many replicates of arrays are required to detect gene expression changes in microarray experiments ? a mixture model approach. *Genome Biol*, 3(5) :research0022, janvier 2002.

[Pan *et al.*, 2002b] W. Pan, J. Lin et C. T. Le. Model-based cluster analysis of microarray gene-expression data. *Genome Biol*, 3(2) :RESEARCH0009, janvier 2002.

[Park *et al.*, 2001] J. Park, M. Lappe et S. A. Teichmann. Mapping protein family interactions : intramolecular and intermolecular protein family interaction repertoires in the pdb and yeast. *J Mol Biol*, 307(3) :929–38, mars 2001.

[Pazos *et al.*, 1997] F. Pazos, M. Helmer-Citterich, G. Ausiello et A. Valencia. Correlated mutations contain information about protein-protein interaction. *J Mol Biol*, 271(4) :511–23, août 1997.

[Pazos et Valencia, 2001] F. Pazos et A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9) :609–14, septembre 2001.

[Pazos et Valencia, 2002] F. Pazos et A. Valencia. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, 47(2) :219–27, mai 2002.

[Pazos et Valencia, 2008] F. Pazos et A. Valencia. Protein co-evolution, co-adaptation and interactions. *EMBO J*, 27(20) :2648–55, octobre 2008.

[Peddada *et al.*, 2003] S. D. Peddada, E. K. Lobenhofer, L. Li, C. A. Afshari, C. R. Weinberg et D. M. Umbach. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics*, 19(7) :834–41, mai 2003.

[Pellegrini *et al.*, 1999] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg et T. O. Yeates. Assigning protein functions by comparative genome analysis : protein phylogenetic profiles. *Proc Natl Acad Sci USA*, 96(8) :4285–8, avril 1999.

[Persico *et al.*, 2005] M. Persico, A. Ceol, C. Gavrila, R. Hoffmann, A. Florio et G. Cesareni. Homomint : an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, 6 Suppl 4 :S21, décembre 2005.

[Peschek, 1996] G. A. Peschek. Structure-function relationships in the dual-function photosynthetic-respiratory electron-transport assembly of cyanobacteria (blue-green algae). *Biochem Soc Trans*, 24(3) :729–33, août 1996.

[Petryszak *et al.*, 2005] R. Petryszak, E. Kretschmann, D. Wieser et R. Apweiler. The predictive power of the clustr database. *Bioinformatics*, 21(18) :3604–9, septembre 2005.

[Picard *et al.*, 2006] F. Picard, J. Daudin et S. Robin. A mixture model for random graphs. *Stat Comput*, 18(2) :173–83, 2006.

[Polaillon *et al.*, 2007] G. Polaillon, L. Vescovo, M. Michaut et J.-C. Aude. Mining biological data using pyramids. *Springer-Verlag*, Studies in Classification, Data Analysis, and Knowledge Organization :397–408, octobre 2007.

[Poncelet *et al.*, 1998] M. Poncelet, C. Cassier-Chauvat, X. Leschelle, H. Bottin et F. Chauvat. Targeted deletion and mutational analysis of the essential (2fe-2s) plant-like ferredoxin in synechocystis pcc6803 by plasmid shuffling. *Mol Microbiol*, 28(4) :813–21, mai 1998.

[Poyatos et Hurst, 2004] J. F. Poyatos et L. D. Hurst. How biologically relevant are interaction-based modules in protein networks ? *Genome Biol*, 5(11) :R93, janvier 2004.

[Przulj, 2007] N. Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2) :e177–83, janvier 2007.

[Przulj *et al.*, 2004] N. Przulj, D. G. Corneil et I. Jurisica. Modeling interactome : scale-free or geometric ? *Bioinformatics*, 20(18) :3508–15, décembre 2004.

[Przulj et Higham, 2006] N. Przulj et D. J. Higham. Modelling protein-protein interaction networks via a stickiness index. *Journal of the Royal Society, Interface / the Royal Society*, 3(10) :711–6, octobre 2006.

[Qi *et al.*, 2005] Y. Qi, J. Klein-Seetharaman et Z. Bar-Joseph. Random forest similarity for protein-protein interaction prediction from multiple sources. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 531–42, janvier 2005.

[Qiu *et al.*, 2006] X. Qiu, Y. Xiao, A. Gordon et A. Yakovlev. Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics*, 7 :50, janvier 2006.

[Quackenbush, 2001] J. Quackenbush. Computational analysis of microarray data. *Nat Rev Genet*, 2(6) :418–27, juin 2001.

[Quackenbush, 2002] J. Quackenbush. Microarray data normalization and transformation. *Nat Genet*, 32 Suppl :496–501, décembre 2002.

[Rain *et al.*, 2001] J. C. Rain, L. Selig, H. D. Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schächter, Y. Chemama, A. Labigne et P. Legrain. The protein-protein interaction map of helicobacter pylori. *Nature*, 409(6817) :211–5, janvier 2001.

[Ramoni *et al.*, 2002] M. F. Ramoni, P. Sebastiani et I. S. Kohane. Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci USA*, 99(14) :9121–6, juillet 2002.

[Reimers et Carey, 2006] M. Reimers et V. J. Carey. Bioconductor : an open source framework for bioinformatics and computational biology. *Meth Enzymol*, 411 :119–34, janvier 2006.

[Reiner *et al.*, 2003] A. Reiner, D. Yekutieli et Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3) :368–75, février 2003.

[Remm *et al.*, 2001] M. Remm, C. E. Storm et E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5) :1041–52, décembre 2001.

[Ritchie *et al.*, 2007] M. E. Ritchie, J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway et G. K. Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20) :2700–7, octobre 2007.

[Rives et Galitski, 2003] A. W. Rives et T. Galitski. Modular organization of cellular networks. *Proc Natl Acad Sci USA*, 100(3) :1128–33, février 2003.

[Rual *et al.*, 2005] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth et M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062) :1173–8, octobre 2005.

[Saebø *et al.*, 2005] P. E. Saebø, S. M. Andersen, J. Myrseth, J. K. Laerdahl et T. Rognes. Paralign : rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res*, 33(Web Server issue) :W535–9, juillet 2005.

[Saeed *et al.*, 2003] A. I. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush et

J. Quackenbush. Tm4 : a free, open-source system for microarray data management and analysis. *BioTechniques*, 34(2) :374–8, février 2003.

[Salwinski *et al.*, 2004] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie et D. Eisenberg. The database of interacting proteins : 2004 update. *Nucleic Acids Res*, 32(Database issue) :D449–51, janvier 2004.

[Sanchez *et al.*, 1999] C. Sanchez, C. Lachaize, F. Janody, B. Bellon, L. Röder, J. Euzenat, F. Rechenmann et B. Jacq. Grasping at molecular interactions and genetic networks in drosophila melanogaster using flynets, an internet database. *Nucleic Acids Research*, 27(1) :89–94, janvier 1999.

[Sásik *et al.*, 2002] R. Sásik, N. Iranfar, T. Hwa et W. F. Loomis. Extracting transcriptional events from temporal gene expression patterns during dictyostelium development. *Bioinformatics*, 18(1) :61–6, janvier 2002.

[Sato *et al.*, 2007] S. Sato, Y. Shimoda, A. Muraki, M. Kohara, Y. Nakamura et S. Tabata. A large-scale protein protein interaction analysis in synechocystis sp. pcc6803. *DNA Res*, novembre 2007.

[Schramm *et al.*, 2007] G. Schramm, M. Zapatka, R. Eils et R. König. Using gene expression data and network topology to detect substantial pathways, clusters and switches during oxygen deprivation of escherichia coli. *BMC Bioinformatics*, 8 :149, janvier 2007.

[Schwikowski *et al.*, 2000] B. Schwikowski, P. Uetz et S. Fields. A network of protein-protein interactions in yeast. *Nat Biotechnol*, 18(12) :1257–61, décembre 2000.

[Scott *et al.*, 2006] J. Scott, T. Ideker, R. M. Karp et R. Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol*, 13(2) :133–44, mars 2006.

[Segal *et al.*, 2003] E. Segal, H. Wang et D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19 Suppl 1 :i264–71, janvier 2003.

[Shannon *et al.*, 2003] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski et T. Ideker. Cytoscape : a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11) :2498–504, novembre 2003.

[Sharp et Li, 1987] P. M. Sharp et W. H. Li. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, 15(3) :1281–95, février 1987.

[Shoemaker et Panchenko, 2007a] B. A. Shoemaker et A. R. Panchenko. Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLoS Comput Biol*, 3(3) :e42, mars 2007.

[Shoemaker et Panchenko, 2007b] B. A. Shoemaker et A. R. Panchenko. Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, 3(4) :e43, avril 2007.

[Smith et Waterman, 1981] T. F. Smith et M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1) :195–7, mars 1981.

[Smyth, 2004] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3 :Article3, janvier 2004.

[Speed, 2003] T. Speed. Statistical analysis of gene expression microarray data. *CRC Press*, 2003.

[Sprinzak et Margalit, 2001] E. Sprinzak et H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4) :681–92, août 2001.

[Stanyon *et al.*, 2004] C. A. Stanyon, G. Liu, B. A. Mangiola, N. Patel, L. Giot, B. Kuang, H. Zhang, J. Zhong et R. L. Finley. A drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biol*, 5(12) :R96, janvier 2004.

[Stark *et al.*, 2006] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz et M. Tyers. Biogrid : a general repository for interaction datasets. *Nucleic Acids Research*, 34(Database issue) :D535–9, janvier 2006.

[Stelzl et Wanker, 2006] U. Stelzl et E. E. Wanker. The value of high quality protein-protein interaction networks for systems biology. *Current opinion in chemical biology*, 10(6) :551–8, décembre 2006.

[Stelzl *et al.*, 2005] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach et E. E. Wanker. A human protein-protein interaction network : a resource for annotating the proteome. *Cell*, 122(6) :957–68, septembre 2005.

[Stohs et Bagchi, 1995] S. J. Stohs et D. Bagchi. Oxidative mechanisms in the toxicity of metal ions. *Free Radic Biol Med*, 18(2) :321–36, février 1995.

[Stumpf *et al.*, 2008] M. P. H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe et C. Wiuf. Estimating the size of the human interactome. *Proc Natl Acad Sci U S A*, 105(19) :6959–64, mai 2008.

[Suleau *et al.*, 2008] A. Suleau, A. Tavenet, R. Ferrari, C. Ducrot, M. Michaut, G. Dieci, O. Lefebvre, C. Conesa et J. Acker. Sub1 regulates genes involved in cell growth and dna damage response, including pol iii-transcribed genes. *(soumis)*, 2008.

[Szilágyi *et al.*, 2005] A. Szilágyi, V. Grimm, A. K. Arakaki et J. Skolnick. Prediction of physical protein-protein interactions. *Physical biology*, 2(2) :S1–16, juin 2005.

[Tamayo *et al.*, 1999] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander et T. R. Golub. Interpreting patterns of gene expression with self-organizing maps : methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA*, 96(6) :2907–12, mars 1999.

[Tarassov *et al.*, 2008] K. Tarassov, V. Messier, C. R. Landry, S. Radinovic, M. M. S. Molina, I. Shames, Y. Malitskaya, J. Vogel, H. Bussey et S. W. Michnick. An in vivo map of the yeast protein interactome. *Science*, 320(5882) :1465–70, juin 2008.

[Tatusov *et al.*, 2003] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin et D. A. Natale. The cog database : an updated version includes eukaryotes. *BMC Bioinformatics*, 4 :41, septembre 2003.

[Tatusov *et al.*, 1997] R. L. Tatusov, E. V. Koonin et D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338) :631–7, octobre 1997.

[Tavazoie *et al.*, 1999] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho et G. M. Church. Systematic determination of genetic network architecture. *Nat Genet*, 22(3) :281–5, juillet 1999.

[Teichmann, 2002] S. A. Teichmann. The constraints protein-protein interactions place on sequence divergence. *J Mol Biol*, 324(3) :399–407, novembre 2002.

[Thomas *et al.*, 2001] J. G. Thomas, J. M. Olson, S. J. Tapscott et L. P. Zhao. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res*, 11(7) :1227–36, juillet 2001.

[Tong *et al.*, 2002] A. H. Y. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. V. Hogue, S. Fields, C. Boone et G. Cesareni. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553) :321–4, janvier 2002.

[Troyanskaya, 2005] O. G. Troyanskaya. Putting microarrays in a context : integrated analysis of diverse biological data. *Brief Bioinformatics*, 6(1) :34–43, mars 2005.

[Troyanskaya *et al.*, 2003] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman et D. Botstein. A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae). *Proc Natl Acad Sci USA*, 100(14) :8348–53, juillet 2003.

[Troyanskaya *et al.*, 2002] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein et R. B. Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11) :1454–61, novembre 2002.

[Tusher *et al.*, 2001] V. G. Tusher, R. Tibshirani et G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98(9) :5116–21, avril 2001.

[Uetz *et al.*, 2000] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields et J. M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770) :623–7, février 2000.

[Ulitsky et Shamir, 2007] I. Ulitsky et R. Shamir. Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology*, janvier 2007.

[Vidal et Legrain, 1999] M. Vidal et P. Legrain. Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Res*, 27(4) :919–29, février 1999.

[von Mering et Bork, 2002] C. von Mering et P. Bork. Teamed up for transcription. *Nature*, 417(6891) :797–8, juin 2002.

[von Mering *et al.*, 2003] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork et B. Snel. String : a database of predicted functional associations between proteins. *Nucleic Acids Res*, 31(1) :258–61, janvier 2003.

[von Mering *et al.*, 2007] C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Krüger, B. Snel et P. Bork. String 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, 35(Database issue) :D358–62, janvier 2007.

[von Mering *et al.*, 2005] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen et P. Bork. String : known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33(Database issue) :D433–7, janvier 2005.

[von Mering *et al.*, 2002] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields et P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887) :399–403, mai 2002.

[Wachi *et al.*, 2005] S. Wachi, K. Yoneda et R. Wu. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, 21(23) :4205–8, décembre 2005.

[Walhout *et al.*, 2000] A. J. Walhout, R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg et M. Vidal. Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, 287(5450) :116–22, janvier 2000.

[Wang et Zhang, 2007] Z. Wang et J. Zhang. In search of the biological significance of modular structures in protein networks. *PLoS Comput Biol*, 3(6) :e107, juin 2007.

[Watts et Strogatz, 1998] D. J. Watts et S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684) :440–2, juin 1998.

[Wheeler *et al.*, 2008] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner et E. Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 36(Database issue) :D13–21, janvier 2008.

[Wojcik *et al.*, 2002] J. Wojcik, I. G. Boneca et P. Legrain. Prediction, assessment and validation of protein interaction maps in bacteria. *J Mol Biol*, 323(4) :763–70, novembre 2002.

[Wojcik et Schächter, 2001] J. Wojcik et V. Schächter. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17 Suppl 1 :S296–305, janvier 2001.

[Wu *et al.*, 2006a] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi et B. Suzek. The universal protein resource (uniprot) : an expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue) :D187–91, janvier 2006.

[Wu *et al.*, 2006b] J. Wu, Z. Hu et C. DeLisi. Gene annotation and network inference by phylogenetic profiling. *BMC Bioinformatics*, 7 :80, janvier 2006.

[Wu *et al.*, 2003] J. Wu, S. Kasif et C. DeLisi. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, 19(12) :1524–30, août 2003.

[Wuchty et Ipsaro, 2007] S. Wuchty et J. J. Ipsaro. A draft of protein interactions in the malaria parasite p. falciparum. *J Proteome Res*, 6(4) :1461–70, avril 2007.

[Xenarios *et al.*, 2001] I. Xenarios, E. Fernandez, L. Salwinski, X. J. Duan, M. J. Thompson, E. M. Marcotte et D. Eisenberg. Dip : The database of interacting proteins : 2001 update. *Nucleic Acids Res*, 29(1) :239–41, janvier 2001.

[Xenarios *et al.*, 2000] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte et D. Eisenberg. Dip : the database of interacting proteins. *Nucleic Acids Res*, 28(1) :289–91, janvier 2000.

[Yamanishi *et al.*, 2004] Y. Yamanishi, J.-P. Vert et M. Kanehisa. Protein network inference from multiple genomic data : a supervised approach. *Bioinformatics*, 20 Suppl 1 :i363–70, août 2004.

[Yang et Yang, 2006] J. J. Yang et M. C. K. Yang. An improved procedure for gene selection from microarray experiments using false discovery rate criterion. *BMC Bioinformatics*, 7 :15, janvier 2006.

[Yang *et al.*, 2006] K. Yang, Z. Cai, J. Li et G. Lin. A stable gene selection in microarray data analysis. *BMC Bioinformatics*, 7 :228, janvier 2006.

[Yeung *et al.*, 2005] K. Y. Yeung, R. E. Bumgarner et A. E. Raftery. Bayesian model averaging : development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, 21(10) :2394–402, mai 2005.

[Yu *et al.*, 2008] H. Yu, P. Braun, M. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. Rual, A. Dricot, A. Vazquez, R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. de Smet, A. Motyl, M. Hudson, J. Park, X. Xin, M. Cusick, T. Moore, C. Boone, M. Snyder, F. Roth, A. Barabasi, J. Tavernier, D. Hill et M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, août 2008.

[Yu *et al.*, 2004a] H. Yu, D. Greenbaum, H. X. Lu, X. Zhu et M. Gerstein. Genomic analysis of essentiality within protein networks. *Trends Genet*, 20(6) :227–31, juin 2004.

[Yu *et al.*, 2007] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov et M. Gerstein. The importance of bottlenecks in protein networks : correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3(4) :e59, avril 2007.

[Yu *et al.*, 2004b] H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J.-D. J. Han, N. Bertin, S. Chung, M. Vidal et M. Gerstein. Annotation transfer between genomes : protein-protein interologs and protein-dna regulogs. *Genome Res*, 14(6) :1107–18, juin 2004.

[Zanzoni *et al.*, 2002] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich et G. Cesareni. Mint : a molecular interaction database. *FEBS Lett*, 513(1) :135–40, février 2002.

[Zhang *et al.*, 2008] Z. Zhang, N. D. Pendse, K. N. Phillips, J. B. Cotner et A. Khodursky. Gene expression patterns of sulfur starvation in synechocystis sp. pcc 6803. *BMC Genomics*, 9 :344, janvier 2008.

[Zotenko *et al.*, 2008] E. Zotenko, J. Mestre, D. P. O'Leary et T. M. Przytycka. Why do hubs in the yeast protein interaction network tend to be essential : reexamining the connection between the network topology and essentiality. *PLoS Comput Biol*, 4(8) :e1000140, août 2008.

Publications

La liste suivante indique les publications ayant été réalisées au cours de ce travail de thèse. Outre les travaux sur la prédiction d'interaction protéine-protéine et la classification mixte présentés dans le manuscrit, j'ai également été impliquée dans deux projets chez la levure portant d'une part sur la localisation du facteur de transcription TFIIS, et d'autre part sur le régulateur *sub1*. Ces travaux sont présentés à la suite des références bibliographiques, sous la forme de publications. Les titres des articles majeurs sont indiqués en gras et le contenu de ces articles est ajouté à la suite de cette liste.

– Journal scientifique :

1. Laetitia Houot, Martin Floutier, Benoit Marteyn, Magali Michaut, Antoine Picciocchi, Pierre Legrain, Jean-Christophe Aude, Corinne Cassier-Chauvat and Franck Chauvat ; *BMC Genomics* 8(1) :350 (2007)
   Cadmium triggers an integrated reprogramming of the metabolism of *Synechocystis* PCC6803, under the control of the Slr1738 regulator [Houot *et al.*, 2007]

2. Magali Michaut, Samuel Kerrien, Luisa Montecchi-Palazzi, Franck Chauvat, Corinne Cassier-Chauvat, Jean-Christophe Aude, Pierre Legrain and Henning Hermjakob ; *Bioinformatics* 24(14) : 1625-1631 (2008)
   InteroPorc : Automated Inference of Highly Conserved Protein Interaction Networks [Michaut *et al.*, 2008d]

3. Yad Ghavi-Helm, Magali Michaut, Joël Acker, Jean-Christophe Aude, Pierre Thuriaux, Michel Werner and Julie Soutourina ; *Genes and Development* 22(14) : 1934-47 (2008)
   Genome-wide location analysis reveals a novel role of TFIIS in RNA polymerase III transcription [Ghavi-Helm *et al.*, 2008]

4. Audrey Suleau, Arounie Tavenet, Roberto Ferrari, Cécile Ducrot, Magali Michaut, Jean-Christophe Aude, Giorgio Dieci, Olivier Lefebvre, Christine Conesa and Joel Acker (soumis)
   Sub1 regulates genes involved in cell growth and DNA damage response, including Pol III-transcribed genes [Suleau *et al.*, 2008]

5. Magali Michaut, Samuel Kerrien, Luisa Montecchi-Palazzi, Franck Chauvat, Corinne Cassier-Chauvat, Jean-Christophe Aude, Pierre Legrain, Henning Hermjakob ; *BMC Bioinformatics* 2008, 9(Suppl 10) :P1 (2008)
   Highlights from the Fourth International Society for Computational Biology (ISCB) Student Council Symposium [Michaut *et al.*, 2008c]

6. Magali Michaut, Henning Hermjakob, Pierre Legrain and Jean-Christophe Aude ; *en préparation*
   Synechocystis protein-protein interaction network [Michaut *et al.*, 2008a]

– Actes de colloques avec comité de lecture :

7. Magali Michaut, Samuel Kerrien, Luisa Montecchi-Palazzi, Franck Chauvat, Corinne Cassier-Chauvat, Jean-Christophe Aude, Pierre Legrain and Henning Hermjakob ; *Proceedings of Journées Ouvertes Biologie Informatique Mathématiques* pages 229-234 (2007)
   Inference and validation of *Synechocystis* protein interaction network using orthology [Michaut *et al.*, 2007]

8. Magali Michaut, Samuel Kerrien, Luisa Montecchi-Palazzi, Franck Chauvat, Corinne Cassier-Chauvat, Jean-Christophe Aude, Pierre Legrain and Henning Hermjakob ; *Proceedings of Journées Ouvertes Biologie Informatique Mathématiques* pages 99-104 (2008)
   Inference of *Synechocystis* protein interaction network [Michaut *et al.*, 2008b]

– Chapitre de livre :

9. Geraldine Polaillon, Laure Vescovo, Magali Michaut, Jean-Christophe Aude ; *Springer-Verlag* pages 397-408 (2007)
   Mining biological data using pyramids, in Studies in Classification, Data Analysis, and Knowledge Organization [Polaillon *et al.*, 2007]

BMC Genomics

BioMed Central

Research article

Cadmium triggers an integrated reprogramming of the metabolism of *Synechocystis* PCC6803, under the control of the Slr1738 regulator

Laetitia Houot[1], Martin Floutier[†1,2], Benoit Marteyn[†1], Magali Michaut[1], Antoine Picciocchi[1], Pierre Legrain[1,2], Jean-Christophe Aude[1], Corinne Cassier-Chauvat[1,2] and Franck Chauvat[*1]

Address: [1]Commissariat à l'Energie Atomique, Institut de Biologie et Tecnologies de Saclay, Service de Biologie Intégrative et Génétique Moléculaire, CEA Saclay F-91191 Gif sur Yvette CEDEX, France and [2]Centre National de la Recherche Scientifique, Unité de Recherche Associée 2096 CEA Saclay, F-91191 Gif sur Yvette CEDEX, France

Email: Laetitia Houot - Laetitia.Houot@childrens.harvard.edu; Martin Floutier - martinfloutier@yahoo.fr; Benoit Marteyn - b.marteyn@imperial.ac.uk; Magali Michaut - magali.michaut@cea.fr; Antoine Picciocchi - antoine_picciocchi@hotmail.com; Pierre Legrain - pierre.legrain@cea.fr; Jean-Christophe Aude - jean-christophe.aude@cea.fr; Corinne Cassier-Chauvat - corinne.cassier-chauvat@cea.fr; Franck Chauvat* - franck.chauvat@cea.fr

* Corresponding author †Equal contributors

Published: 2 October 2007

BMC Genomics 2007, **8**:350 doi:10.1186/1471-2164-8-350

Received: 8 June 2007
Accepted: 2 October 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/350

Abstract

Background: Cadmium is a persistent pollutant that threatens most biological organisms, including cyanobacteria that support a large part of the biosphere. Using a multifaceted approach, we have investigated the global responses to Cd and other relevant stresses (H_2O_2 and Fe) in the model cyanobacterium *Synechocystis* PCC6803.

Results: We found that cells respond to the Cd stress in a two main temporal phases process. In the "early" phase cells mainly limit Cd entry through the negative and positive regulation of numerous genes operating in metal uptake and export, respectively. As time proceeds, the number of responsive genes increases. In this "massive" phase, Cd downregulates most genes operating in (i) photosynthesis (PS) that normally provides ATP and NADPH; (ii) assimilation of carbon, nitrogen and sulfur that requires ATP and NAD(P)H; and (iii) translation machinery, a major consumer of ATP and nutrients. Simultaneously, many genes are upregulated, such as those involved in Fe acquisition, stress tolerance, and protein degradation (crucial to nutrients recycling). The most striking common effect of Cd and H_2O_2 is the disturbance of both light tolerance and Fe homeostasis, which appeared to be interdependent. Our results indicate that cells challenged with H_2O_2 or Cd use different strategies for the same purpose of supplying Fe atoms to Fe-requiring metalloenzymes and the SUF machinery, which synthesizes or repairs Fe-S centers. Cd-stressed cells preferentially breakdown their Fe-rich PS machinery, whereas H_2O_2-challenged cells preferentially accelerate the intake of Fe atoms from the medium.

Conclusion: We view the responses to Cd as an integrated "Yin Yang" reprogramming of the whole metabolism, we found to be controlled by the Slr1738 regulator. As the Yin process, the ATP- and nutrients-sparing downregulation of anabolism limits the poisoning incorporation of Cd into metalloenzymes. As the compensatory Yang process, the PS breakdown liberates nutrient assimilates for the synthesis of Cd-tolerance proteins, among which we found the Slr0946 arsenate reductase enzyme.

Background

Photosynthetic organisms that support much of the life on Earth, in using solar energy to renew the oxygenic atmosphere and make up organic assimilates essential to the food chain [1,2] are frequently challenged with toxic reactive oxygen species (ROS) generated by respiration and photosynthesis [3], and toxic metals that constitute persistent pollutants because they cannot be degraded. One of them, Cadmium (Cd), is very abundant in the environment as it is often combined with sulfur in Earth's crust, and it is also intensively spread out as (i) a by-product of zinc mining, (ii) the burning of fossil fuel, (iii) the dispersal of sewage sludge and phosphate fertilizers, and (vi) the manufacturing of paints, batteries and screens [4]. Subsequently, Cd can be transferred to the food chain, and bio-accumulated in human where it has a half-life greater than 20 years [5] and causes various diseases by as yet unclear processes [6]. Even metals that are essential to enzyme activity, such as zinc and iron [7,8], can become toxic when occurring in excess. This toxicity is likely due to the poisoning replacement of the cognate metal cofactor of diverse metalloenzymes, a phenomenon sometimes leading to oxidative stress [9].

Cyanobacteria, the most abundant photosynthetic organisms on Earth [10], are attractive models to investigate the interrelations between metal toxicity and oxidative stress, because they perform the two metal-requiring [8] ROS-generating processes [3], photosynthesis and respiration, in the same membrane system [11]. Furthermore, cyanobacteria share a wide range of genes in common with plants [12], in agreement with they being the likely ancestor of chloroplast [13]. Thus, lessons learned from stress responses in cyanobacteria will also greatly facilitate the understanding of how plant cells face environmental challenges. This is important, as Cd has been reported to be toxic to plants by as yet unknown processes that may [14] or may not [15] impair photosynthesis. Moreover, cyanobacteria are also suitable for biosensor and/or bioremediation applications [16,17].

Using the model cyanobacterium *Synechocystis* PCC6803 that possesses a small genome [18] fully sequenced, and easily manipulable with replicating plasmids [19-21], we have analyzed the global responses of photosynthetic cells challenged with Cd, H_2O_2 (the paradigm ROS agent) or drastic changes of availability of either Fe or Zn, through (i) DNA microarrays; (ii) absorption spectroscopy; (iii) oxygen evolution; (iv) Western blot; (v) targeted gene inactivation and (vi) assays of cell fitness. We show that Cd triggers a "Yin Yang" integrated reorganization of the cyanobacterial metabolism, under the control of the Slr1738 regulator. The "Yin" ATP-sparing downregulation of cell metabolism likely limits Cd uptake and poisoning incorporation in place of the cognate metal cofactor of metalloenzymes. The compensatory "Yang" breakdown of the photosynthetic machinery that impairs ATP production, liberates nutrient assimilates that become available for the synthesis of Cd-toxicity protecting enzymes, among which we found the Slr0946 arsenate reductase.

Results

Transcriptional regulations elicited by Cd are slower and more sustained than those triggered by H_2O_2

The transcriptome approach was used to characterize the kinetics of global changes in *Synechocystis* PCC6803 (*Synechocystis*) gene expression elicited by noxious agents, which were continuously applied to the cells to mimic the persistent character of stresses encountered in Nature. Exponentially growing cells were exposed to $CdSO_4$ (50 μM) or H_2O_2 (3 mM) for increasing periods of time that triggered a wide range of changes in cellular viability (from 100% to less than 10%) and regulation (number of responsive genes and magnitude of changes in expression), as required for a thorough investigation of stress responses (Table 1). For each time point, total RNA were isolated from stressed and unstressed cells, reverse-transcribed, differentially labeled (dye swapped), hybridized together (stressed versus unstressed samples) and analyzed with DNA glass microarrays (two slides per each time point), as described in Methods. Our data (Table 1

Table 1: Influence of Stress on the Viability and Global Transcription Profile

Treatment	**Cadmium**										**Hydrogen peroxide**					**Fe depletion**		**Fe excess**		**Zn excess**	
[C]	*50 μM*										*3 mM*					*2-0*	*1-0*	*16 mM*		*776 μM*	
Time (min)	**15**	**30**	**60**	**75**	**90**	**180**	**300**	**300'**	**360**	**960**	**15**	**30**	**180**	**300**	**420**	**2,800**	**2,800**	**240**	**360**	**30**	**240**
% survival	100	98	90	88	85	70	57	59	43	<10	100	98	64	45	19	55	43	80	15	99	70
Induced genes	22	88	46	52	293	299	310	451	283	51	447	408	170	68	75	106	154	42	34	38	245
Repressed gene	8	17	10	13	250	315	328	439	310	26	490	478	92	12	55	104	109	98	100	22	221

Wild type cells were challenged for the indicated durations prior to survival and transcriptome analyses (Methods). In the case of the Fe starvation stress, the switch in Fe concentration is indicated as *1-0* (standing for 1 μM then 0 μM) or *2-0* (standing for 2 μM then 0 μM). The time point 300 min of the kinetic of Cd responses was repeated twice (independent biological repeats, columns 300 and 300') with the two different versions of the Cyanochip DNA microarrays, leading to very reproducible results (see Additional files 2 and 4).
Genes were considered differentially regulated whenever their level of expression was changed at least 1.9 fold.

and see Additional files 1, 2, 3, 4) showed that the Cd-elicited regulation could be divided in two main temporal phases. The early phase was moderate since only 151 genes responded to Cd during the first 60 min of treatment, and the changes were mostly up-regulation. The second phase occurring between 90 to 360 min of treatment was massive, with about 1,222 responsive genes equally distributed between up- and down-regulated genes. The relevance of the "early" and "massive" phases of Cd responses was verified by performing an independent biological repeat of both the time points 60 min (early phase) and 300 min (massive phase), and using an appropriate statistical test (Methods) to analyze all data (see Additional files 2, 3). Indeed, a large number of the Cd-regulated genes (791) appeared to be differentially expressed between the two temporal phases of responses.

By contrast, the transcriptional responses to H2O2 (3 mM) were faster and briefer than those to Cd (Table 1 and see Additional files 2, 3, 4). The massive phase of H2O2-mediated regulation encompassed the time points 15 min and 30 min (1,300 genes controlled, equally distributed between up- and down-regulation), while the late phase occurred between 180 min and 420 min (344 genes controlled, mostly positively), a time period in which most fast-responsive genes had returned to normal expression level (see Additional file 3).

Cadmium antagonistically controls the genes operating in protein synthesis (downregulation), and protein maturation and degradation (upregulation)

Among the earliest responses to Cd (noticeable within the first 30 min of exposure) was the upregulation (see Additional file 4 panel A) of chaperones and proteases genes, the number of which increased during the massive phase of responses (after 60 min.). This regulation was accompanied with the downregulation of most ribosomal proteins genes (noticeable at 90 min, see Additional file 4 pannel A), and, comforting our data, we noticed that operonic genes were co-regulated. By contrast, most aminoacyl-tRNA synthetases genes were unaffected by Cd (see Additional files 2 and 6). Considering the normal level of expression [22] and the response to Cd (this study) of aminoacyl-tRNA synthetases genes (moderate expression, unresponsive to Cd) and ribosomal proteins genes (strong expression, turned down by Cd), we think that Cd-challenged cells preferentially downregulate those genes whose expression represents a metabolic burden. This interpretation is comforted by the findings that photosynthesis genes normally expressed to a high level [22] were also turned down by Cd (see below).

Similarly, H_2O_2 downregulated ribosomal protein genes (see Additional file 4, panel A), and did not affect aminoacyl-tRNA synthetase genes (see Additional file 3).

Also interestingly, we found that Zn excess partly mimics the Cd-mediated control of genes involved in protein folding and turnover (upregulation) or protein synthesis (downregulation), which were little affected by Fe availability (Table 1 and see Additional file 4).

Cd and to a lesser extent H_2O_2 downregulate photosynthesis genes

A very important target of Cd toxicity was the photosynthesis (PS) machinery that uses several electron-transfer complexes, the photosystemII (PSII) and its phycobilisomes (PBS) antennae, the cytochrome b6/f, the photosystemI (PSI) and the ATPase, to produce ATP and NADPH [8] required for the assimilation of inorganic nutrients. In addition, the PS activity can also generate toxic reactive oxygen species (ROS) in case of light excess [23]. Most of these PS genes were downregulated after 75–90 min. of Cd challenge (see Additional file 4 pannel B). The validity of these data was substantiated by the observed co-regulation of the following operons *psaAB*, *psbCD1*, *apcABC* and *atpIHGFDACBE*. In addition, we also observed the well-known [24] antagonistic iron regulation of the ssl0020 ferredoxin gene (repressed by Fe starvation) and *isiB* flavodoxin gene (induced by Fe limitation). Furthermore, we have verified the Cd-elicited downregulation of the *psaC* gene at the level of protein abundance (Fig. 1). Also consistent with the Cd-elicited downregulation of PS genes, we found (see Additional file 4 pannel B) that Cd (i) repressed most pigment synthesis genes, namely: *hemA*, *hemL*, *hemB*, *hemE*, *hemF*, *hemN*, *chlN*, *chlB*, *chlL*, *por*, *ho1*, *ho2*, *cbiX*, *crtH*, *crtR*, *crtD*-homolog and *alg*-homolog, and (ii) induced the *nblA* operonic genes that operate in PBS degradation [25].

The global downregulation of PS and pigment synthesis genes has been observed in cells challenged by a high light stress [26-28] and, very interestingly, we noticed that many of the high light-inducible genes [25,29,30] were also upregulated by Cd (see Additional file 4 panel B) namely: *hliB*, *hliC*, *isiA* and *nblA*. Collectively, these findings suggested that Cd-exposed cells become light sensitive, an interpretation we validated through growth assays (Fig. 2C).

The aspects of Cd toxicity resembling light stress are presumably due to oxidative stress since they could be elicited by H_2O_2 too (see Additional file 4 panel B). These common responses included the downregulation of all ATPase genes and several PBS genes (*apcA*, *apcB*, *apcC*, *apcE* and *apcF*), as well as the concomitant upregulation of many high light-inducible genes (*hliC*, *isiA* and *nblA*). Unsurprisingly, H_2O_2 mimicked high light stress that generates ROS [3] more efficiently than Cd. Indeed (see Additional file 4 panel B), H_2O_2 upregulated several high light-inducible genes unaffected by Cd, namely: *hliA*, *hliD*, *ctpA*

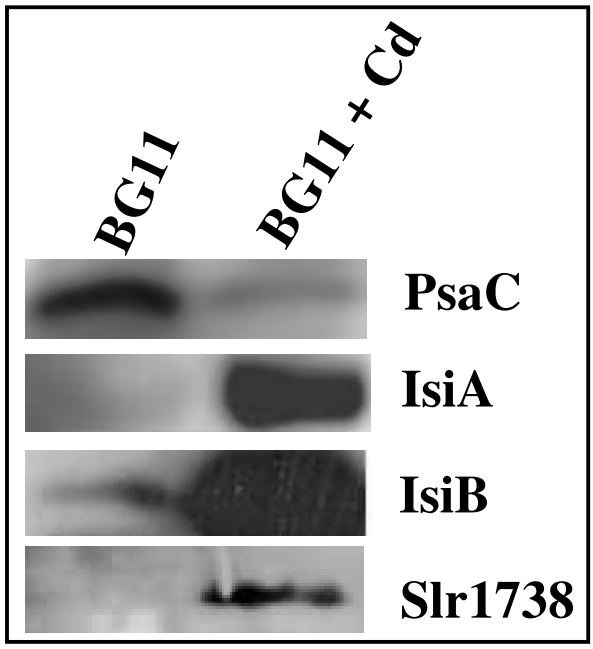


Figure 1
Influence of cadmium on the abundance of selected proteins. Cells were incubated for the indicated durations on solid media with or without (0 for untreated control) $CdSO_4$ (50 μM, 360 min.) prior to disruption. 5 μg of crude cell extracts were analyzed by Western blottings (Methods), using the antibodies directed against the indicated proteins.

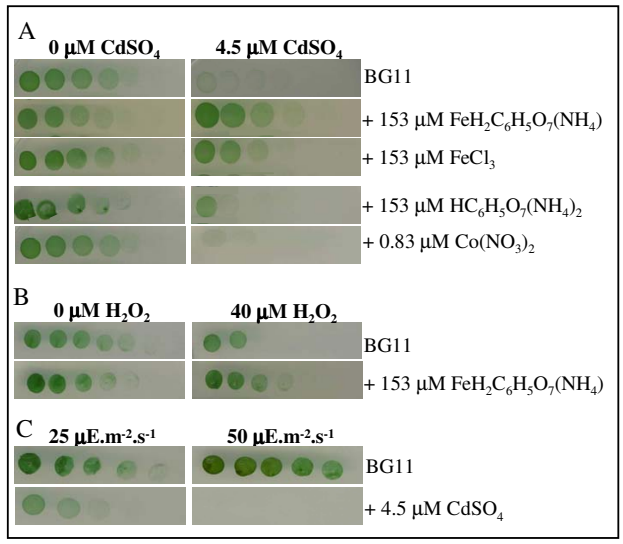


Figure 2
Effect of metal, hydrogen peroxide and light fluence on cellular growth. Four fold serial dilutions of mid log phase liquid cultures were spotted onto BG11 plates with or without the indicated agents, incubated for 4–5 days and scanned (Methods). Panel A: Typical influence of iron ($(NH_4)FeH_2C_6H_5O_7$ or $FeCl_3$), cobalt ($Co(NO_3)_2$) and citrate ($(NH_4)_2H_2C_6H_5O_7$) on the toxicity of cadmium ($CdSO_4$). Panel B: Typical influence of iron ($(NH_4)FeH_2C_6H_5O_7$) on the toxicity of hydrogen peroxide (H_2O_2). Panel C: Influence of the light fluence on the toxicity of $CdSO_4$. These experiments were repeated three times.

and *ftsH* (slr0228 and slr1604) the protease genes involved in the high-light induced turnover of the D1 protein of PSII [31]. Also interestingly, many PS genes downregulated by Cd were actually upregulated by H_2O_2, namely: PSII (*psbB*, *psbJ*, *psbV* and *psbU*), PSI (*psaF*, *psaJ*, *psaD*, *psaI*, *psaM*) and PBS (*cpcC1*, *cpcC2* and *cpcD*) (see Additional file 4 panel B).

In agreement with the upregulation of many genes induced by high light that triggers oxidative stress, we found (see Additional file 4 panel D) that Cd and/or H_2O_2 upregulated many anti-oxidant genes encoding thioredoxin reductase (*trxR*), thioredoxins (*trxA*, *trxM*), glutathione peroxidase (*gpx*), glutaredoxins (*grx*), glyoxalases (*glo*) and peroxidases (*gpx* and *ahpc*).

Similarly to Cd, Zn downregulated numerous PS genes (PBS, PSII, PSI and pigment synthesis, but not ATPase genes), and upregulated genes involved in protein turnover and tolerance to light/oxidative stress (see Additional file 4 panel B). By contrast, Fe controlled a few PS genes.

Also interestingly, the differential regulation of the cytochrome b6/f genes (TableS4B), encoding the predominant (*petC1*) or accessory (*petC2* and *petC3*) Rieske iron-sulfur proteins [32], strongly suggests that alternative b6/f complexes are synthesized in response to changing environmental conditions.

Spectroscopic confirmation that Cd elicits a more intense decline of the photosynthetic machinery than H_2O_2

That Cd- and Zn-excess turned down most photosynthesis genes and simultaneously upregulated protein degradation genes, suggested to us that these stresses decrease the abundance of the PS machinery. By comparison, we anticipated H_2O_2 to elicit a lower decline of the PS apparatus, in downregulating a smaller number of PS genes (see above). These predictions were all validated by another global method i.e. absorption-spectroscopy, which showed that the cellular content of colored PS pigments was decreased strongly in response to Cd- and Zn-stresses (Fig. 3A and 3B) and weakly in response to H_2O_2 (Fig. 3D). As control experiments, we have verified that excess of Fe (with little influence PS-gene expression see above) or cobalt did not alter pigments content (Fig. 3C and 3F).

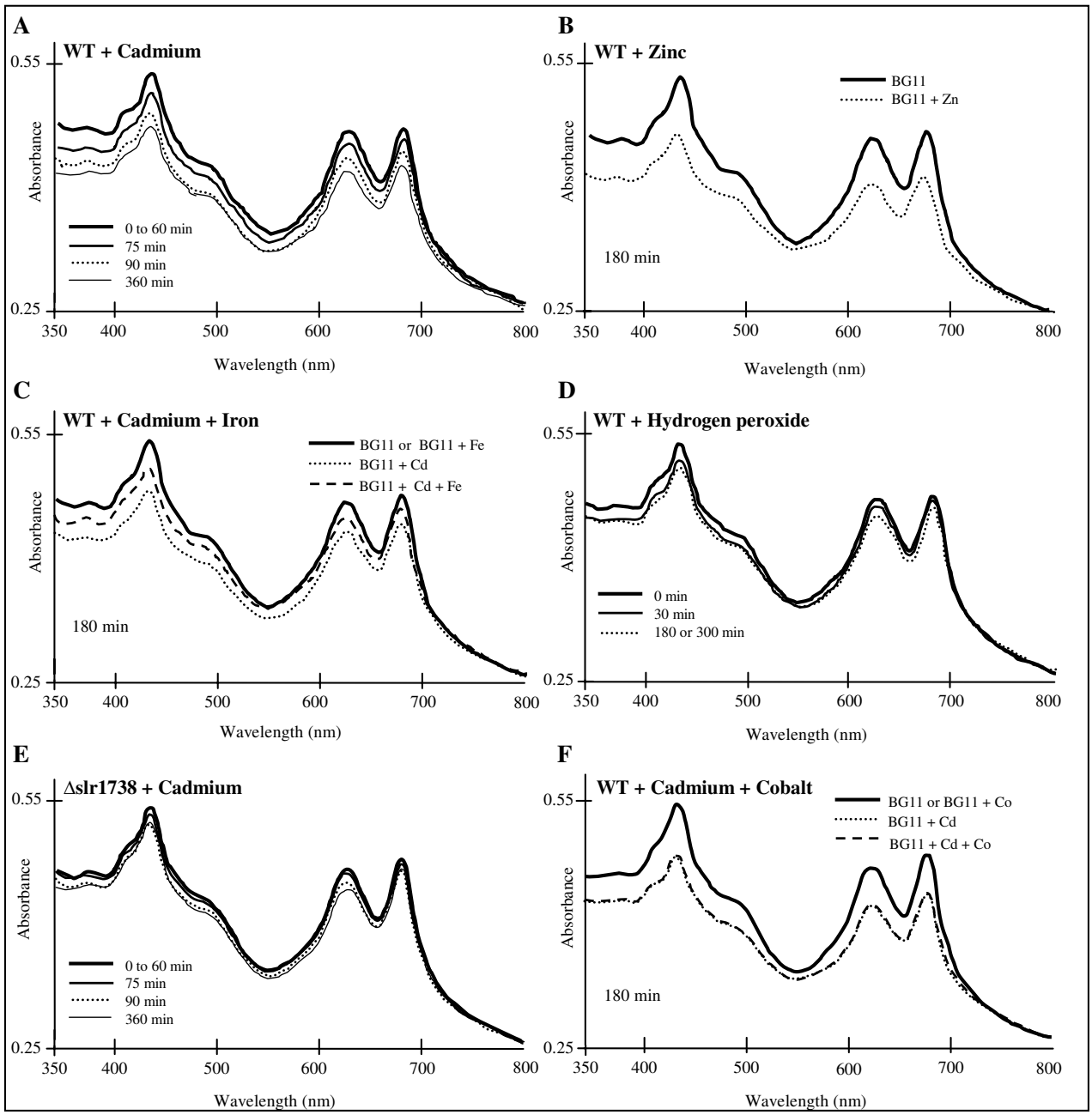


Figure 3

Influence of metals and H_2O_2 on the abundance of photosynthetic pigments. Typical absorption spectra of the wild type (WT) or slr1738 null-mutant (Δslr1738) cells following incubation for the indicated durations on solid BG11 medium with or without H_2O_2 (3 mM), $CdSO_4$ (50 μM), $Co(NO_3)_2$ (350 μM), $(NH_4)FeH_2C_6H_5O_7$ (350 μM) or $ZnSO_4$ (350 μM or 776 μM). The spectra (normalized to light scattering at 800 nm) are displayed in panels A to F. These experiments were repeated three to five times.

Oxygen evolution confirmation that Cd impairs photosynthesis

To further demonstrate that Cd impairs photosynthesis we measured the rate of the whole photosynthetic electron transport (from H_2O_2 to CO_2) of intact cells incubated with or without 50 μM Cd. As expected, the oxygen-evolving activity of Cd-treated cells was strongly decreased

(2.5- and 7-fold after 3- and 6-h, respectively) as compared to that of untreated cells.

Cd and H_2O_2 likely disturb metal homeostasis

Cd rapidly and continuously altered expression of numerous metal transport genes, indicating that it disturbs metal homeostasis (see Additional file 4 panel C). For instance, all members of the nine genes cluster involved in the tolerance to Ni (*nrsBACD* operon), Co (*coaRT* divergon, sll0794 and slr0797) and Zn (*ziaBR* operon and *ziaA* export ATPase) were upregulated by Cd. As one of the numerous findings attesting the relevance of our transcriptome data we observed (see Additional file 4 panel C) that Zn controlled the genes *znuA* (slr2043, Zn uptake, downregulation) and *ziaA* (Zn export, upregulation), as expected [24,33,34]. That Cd regulated both *znuA* (negatively) and *ziaA* (positively), whose product is homologous to the Cd-transporting ATPase CadA [35], suggesting that Cd might be transported via Zn transport systems. Cd also controlled the *corR-corT* divergon operating in Co efflux, as well as the *cbi* cluster and the *cbiX* gene involved in the biosynthesis of cobalamin the Co-dependent vitamin B12 [36]. These data suggest that Cd disturbs Co homeostasis and utilization.

A large part of the numerous genes (more than 20) dedicated to Fe acquisition (*feoB*, *fec*, *fhu* and *fut*) were found to be positively regulated by Fe starvation and turned down by Fe excess (see Additional file 4 pannel B), in agreement with previous Northern blot data [37]. Again, attesting the relevance of our data, we also observed the Fe starvation-mediated control (see Additional file 4 panel D) of the *isiAB* operon (upregulation) and the *fed1* genes (downregulation), as expected [38].

H_2O_2 upregulated all Fe acquisition genes (see Additional file 4 pannel C), as well as (see Additional file 4 panel D) the *suf* genes involved in iron-sulfur cluster biogenesis [39]. These findings are reminiscent to what occurs in *E. coli* where oxidative-stressed cells induce Fe uptake and *suf* genes to accelerate the supply of Fe atoms for the reconstitution of damaged iron-sulfur clusters, in a process leaving no free Fe atoms available for the toxic Fenton chemistry [40-42]. Interestingly, Cd upregulated antioxidant and *suf* genes, as well as half the number of the Fe-uptake genes (see Additional file 4 panels C and D). These findings suggest that Cd damages Fe-S centers, and that the extra Fe atoms required for their repair might be provided not only by the presumably moderate increase in Fe uptake, but also by the breakdown of the Fe-rich photosynthetic machinery (Fig. 3A) that contains 21–23 iron atoms per PS unit [43].

Iron availability controls the Cd-elicited decline of cell viability and PS machinery

The above-mentioned data led us to predict that Fe availability can influence not only cell tolerance to H_2O_2 and Cd, but also the Cd-elicited decline of the PS machinery. As anticipated, we found that the addition of Fe in the medium at the onset of the stresses increased cell resistance to H_2O_2 and Cd (Fig. 2A and 2B), and prevented the Cd-elicited decline of the PS machinery (Fig. 3C). As a negative control, we verified that cobalt (Co) was unable to mimic these Fe-mediated protection effects (Fig. 2A and Fig. 3F).

The Slr0946 arsenate reductase contributes to cadmium tolerance

We noticed that the *arsBHC* tricistronic operon (slr0944 to slr0946) operating in arsenic resistance [44,45] was rapidly and continuously upregulated by Cd (see Additional file 4 panel C). To confirm that the ArsC arsenate-reductase enzyme is a key factor in the tolerance to cadmium, we have deleted the *arsC* gene (Methods), and found the corresponding fully-viable *arsC* null mutant to be more sensitive to Cd than the WT strain (Fig. 4).

Cd and H_2O_2 downregulate carbon metabolism genes, many of which encode ATP-requiring enzymes

Most CO_2 concentrating mechanism (CCM) genes for the acquisition and assimilation of inorganic carbon (Ci) [46] were downregulated by Cd (see Additional file 4 pannel E), namely: (i) the *ndhF3*, *ndhD3* and *cupA* tricistronic operon (CO_2 uptake, NDH-I_3 system); (ii) the *ndhF4* and *ndhD4* operon and the *cupB* gene (CO_2 uptake, NDH-I_4 system); (iii) the *cmpABCD* operon (HCO_3^- transporter); (iv) the *sbtAB* operon (HCO_3^- transporter); (v) the *cca* gene (carbonic anhydrase); (vi) the carboxysome genes *ccmK4* and *ccmK-N* operon; (vii) the *prk* gene (phosphoribulokinase) and (viii) the *ppc* gene (phosphoenol pyruvate carboxylase). These results, together with the constitutive expression of the low-Ci inducible gene *ndhR* encoding the Ci-assimilation regulator [47], indicate that Cd challenged cells are not suffering from Ci starvation. Similarly, most carbon metabolism genes were turned down by Cd (see Additional file 4 panel E), namely: (i) *gpmA* (phosphoglycerate mutase); (ii) *eno* (enolase); (iii) *pyk2* (pyruvatekinase); (iv) *pgk* (phosphoglycerate kinase); (v) *gap2* (G3P-dehydrogenase); (vi) *glgC* (glucose-1-phosphateadenylyltransferase gene) involved in glycogen synthesis; (vii) *pgm* (phosphoglucomutase); (viii) *pfkA* (phosphofructokinaseI); (ix) *fbpI* (fructose1,6-biphosphataseI); (x) *fbaA* (fructose biphosphatealdolaseII); (xi) *pdhABCD* (pyruvate dehydrogenase); (xii) *icd* (isocitrate dehydrogenase).

Collectively, these data suggest that Cd downregulates citrate synthesis and conversion into 2-oxoglutarate that

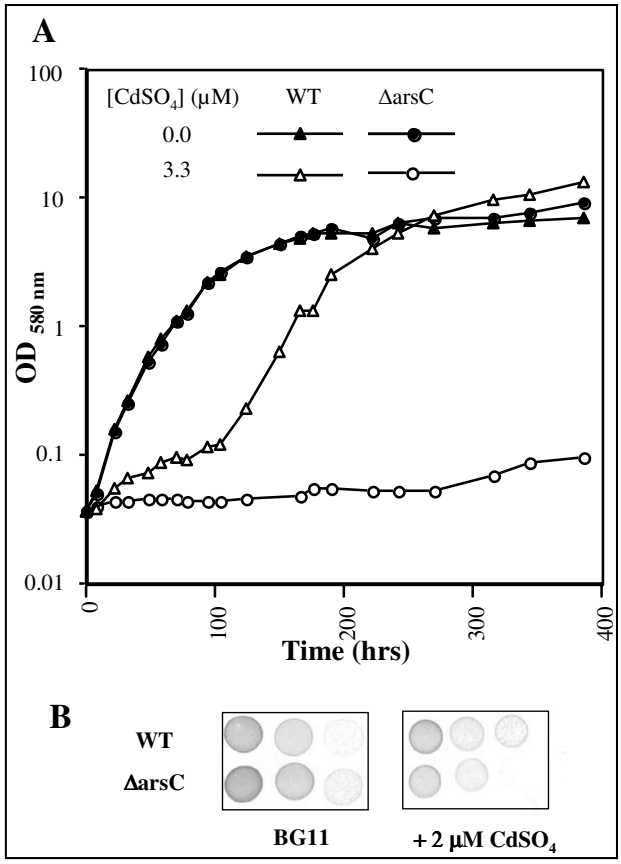


Figure 4
Influence of cadmium on the growth of the wild type strain and *ΔarsC* mutant. Typical growth of the wild type (WT, triangles) and *ΔarsC* mutant (circles) cultivated in liquid (PanelA) or solid (PanelB) BG11 media with or without $CdSO_4$ (at the indicated concentration). These experiments were repeated three times.

connects carbon and nitrogen assimilation pathways. This prediction was substantiated by the Cd-mediated downregulation of numerous genes operating in nitrogen metabolism (see below).

Very interestingly, we noticed that many of the carbon metabolism genes downregulated by Cd code for ATP-consuming enzymes such as *cmpABCD*, *fbpI*, *pgk*, *pfkA*, *prk* and *pyk1*. This negative regulation can be viewed as a part of a global ATP-sparing response to the Cd stress (See below).

Similarly to Cd, H_2O_2 downregulated many Ci acquisition genes (see Additional file 4 panel E): *ccm*, *cmp* and *sbt*, as well and numerous carbon metabolism genes: *pgk*, *gpmB* (slr1124 and slr1945), *eno*, *fbpI*, *carA*, *carB*, *pdhA*, *pdhB*, *pdhC* and *pdhD*. By contrast, Fe and Zn downregulated a few Ci acquisition genes such as *cmp* and *sbt* (Fe) and *ccm* (Zn), and had very little influence on carbon metabolism genes.

Cd and H_2O_2 downregulate nitrogen metabolism genes, many of which encode ATP-consuming enzymes

Most genes for the acquisition and assimilation of nitrogen were negatively regulated by both Cd and H_2O_2 (see Additional file 4 panel F), namely: (i) *amt1* and *amt2* (ATP-requiring ammonium permease); (ii) *nrtABCD* operon (ATP-requiring uptake of nitrate); (iii) *urtABC* (ATP-requiring uptake of urea); (iv) *narB* (nitrate reductase); (v) *nirA* (nitrite reductase); (vi) *glnA* and *glnN* (the two ATP-requiring glutamine synthase); (vii) *murI* (peptidoglycan synthesis); (viii) *argB* (ATP-dependent N-acetylglutamate kinase for arginine synthesis), (ix) *cphA* (ATP-requiring [48] cyanophycin synthetase); (x) *hemA*, *hemF* and *hemL* (synthesis of PS pigments, see Additional file 4 panel B). Consistent with the downregulation of the glutamine synthase *glnA* gene we found that both Cd and H_2O_2 upregulate the *gifA* and *gifB* genes, which code for an inhibitor of GlnA activity [49]. In addition a few related genes were specifically downregulated by either Cd (*hemE* and *hemN*), or H_2O_2. The latter were the following: (i) *ureA* and *ureF* (urease); (ii) *carAB* (ATP-requiring carbamoyl phosphate synthase); (iii) *glsF* (ferredoxin-dependent glutamate synthase); (iv) *arG* (ATP-requiring argininosuccinate synthase for arginine synthesis); (v) *argH* (argininosuccinate lyase); (vi) *cphA* (ATP-requiring [48] cyanophycin synthetase); and (vii): *proA* (ATP-dependent gamma-glutamyl phosphate reductase for proline synthesis).

Together, our data strongly show that *Synechocystis* challenged with H_2O_2 or Cd downregulates numerous key genes encoding ATP-consuming enzymes involved in nitrogen acquisition and metabolisms. This finding is consistent with the above-mentioned negative regulation of ATP-requiring mechanisms for carbon assimilation and metabolism, and global protein synthesis (See above). We view these downregulations as an ATP-sparing process aimed at compensating the decline in ATP production caused by the negative regulation of ATPase and photosynthesis genes.

Fe (but not Zn) regulated numerous N acquisition and assimilation genes, suggesting that Fe homeostasis and nitrogen assimilation are intrinsically connected.

Cd and H_2O_2 downregulate the two sulfur assimilation genes encoding ATP-dependent enzymes

Very interestingly, the genes *met3* (sulfate adenylyltransferase) and *cysC* (adenylylsulfate kinase) encoding the two ATP-requiring enzymes of the cysteine-synthesis pathway appeared to be downregulated by both Cd and H_2O_2 (see

Additional files 2, 3, 4). These data substantiate the Cd- and H_2O_2-elicited downregulation of ATP-consuming metabolic enzymes mentioned above.

Prominent role of the Slr1738 regulator in the transcriptional responses and survival to Cd

To demonstrate that the Cd-elicited breakdown of the photosynthetic (PS) machinery (Fig. 3A) is a direct physiological response rather than a side effect of cell damage, we searched for a regulator controlling this breakdown process with the view that its inactivation in interfering with the PS decline should decrease the level of tolerance to Cd. Hence, we became interested in the slr1738 transcription regulator gene because it is upregulated by Cd (Fig. 1 and see Additional file 4 panel C), a finding which suggests that Slr1738 might be involved in the responses to Cd. We have deleted the slr1738 gene (see Methods) and found the corresponding slr1738 null-mutant (Δslr1738) to be fully viable in standard growth conditions (Fig. 5), in agreement with the small number of genes with an altered level of expression (23, data not shown). As expected, the Δslr1738 mutant was found to be less resistant to Cd than the WT strain (Fig. 5B and 5D), indicating that Slr1738 mediates some of the Cd-elicited regulations. To characterize the Slr1738-mediated responses to Cd, we used DNA microarrays to identify the genes whose transcript abundance in Cd-treated cells differed at least twofold between the Δslr1738 mutant and the WT strain. As expected, we found that the Cd-elicited downregulation of PS genes (PSII large subunits, PBS, pigments synthesis, ATPases, cytochrome b6/f complex) and the simultaneous upregulation of the *nblA* genes operating in phycobilisomes (PBS) degradation were all impaired in the slr1738 null mutant (see Additional file 4 panel B). These findings were confirmed through absorption-spectroscopy analyses of the cellular content of PS proteins. As expected, the Cd-elicited decline of pigmented proteins was truly lower in the Δslr1738 mutant than in the WT strain (compare Fig. 3E with Fig. 3A). The Cd-elicited downregulation of the genes coding for ribosomal proteins was also impaired in the Δslr1738 mutant (see Additional file 4 panel B). Slr1738 was also found to contribute to the Cd-mediated upregulation of the *suf* genes involved in Fe-S cluster assembly and repair (see Additional file 4 panels C and D), and the *ars* genes operating in tolerance to arsenic and cadmium (Fig. 4).

By contrast, Slr1738 is likely not involved in the Cd-mediated regulation of carbon metabolism (see Additional file 4 panel E), indicating that other Cd response regulator(s) remain to be identified.

Finally, the Δslr1738 mutant appeared to be more resistant to H_2O_2 and paraquat than the WT strain (Fig. 5A and 5C). This phenotype, unnoticed by previous workers [50,51], is consistent with the increased level of expression of various antioxidant genes (see Additional file 4) such as the sll1621 peroxiredoxin gene [50-52] and the sequence homology between Slr1738 and the *B. subtilis* PerR (peroxide) regulator whose inactivation increases the resistance to oxidative stress [53].

Discussion

Photosynthetic organisms that support much of the biosphere are increasingly challenged by heavy metals, which are persistent in the environment since they cannot be degraded. Using the model cyanobacterium *Synechocystis* PCC6803 as the host, we have performed a thorough multidisciplinary analysis of the global responses of photosynthetic cells to cadmium (Cd), as well as to hydrogen peroxide (H_2O_2) and noxious concentrations of the essential metals iron and zinc because the disturbance of metal homeostasis can generate oxidative stress [9]. The presently reported data on the global responses to both the Cd and Zn stresses are novel. In the case of the H_2O_2 and Fe stresses our data confirmed and extended those obtained previously after less-extensive investigations. In the case of the H_2O_2 stress, the novelty of our report is that our temporal analysis made it possible to discriminate between early and late responses, unlike the previously performed single time-point analysis [51]. Concerning the Fe-starvation stress, we choose to study the responses to a continuous Fe limitation because stresses occurring in Nature are durable, whereas Singh and co-workers studied cells recovering from a transient Fe deficiency [38]. This presumably explains the fact that we observed the induction of a larger number of Fe acquisition genes than the previous workers. Furthermore, unlike previous workers we also studied the response to Fe excess to identify those genes oppositely responding to excess and deficiency of Fe, which are therefore likely responding to Fe *per se* rather than to another indirect stress signal.

The presently reported occurrence of two temporal phases in the responses to Cd and H_2O_2, with a stress-specific timing, emphasizes on the value of kinetic analyses of stress responses especially when they are to be compared. Indeed, one cannot assume that two stresses of equal duration and toxicity have the same biological effects, in term of gene regulation. This important fact is illustrated by our findings that cells challenged with Cd (50 μM) or H_2O_2 (3 mM) for the same period of time (i.e. 30 min.) leading to equal lethality (2%) displayed massive responses to H_2O_2 but only moderate (early) responses to Cd in term of the number of the genes regulated (see Additional files 2, 3, 4).

The biological significance of the genome-wide transcriptional responses to all presently tested stresses (Table 1, and see Additional file 1, 2, 3, 4, 5, 6) was validated using

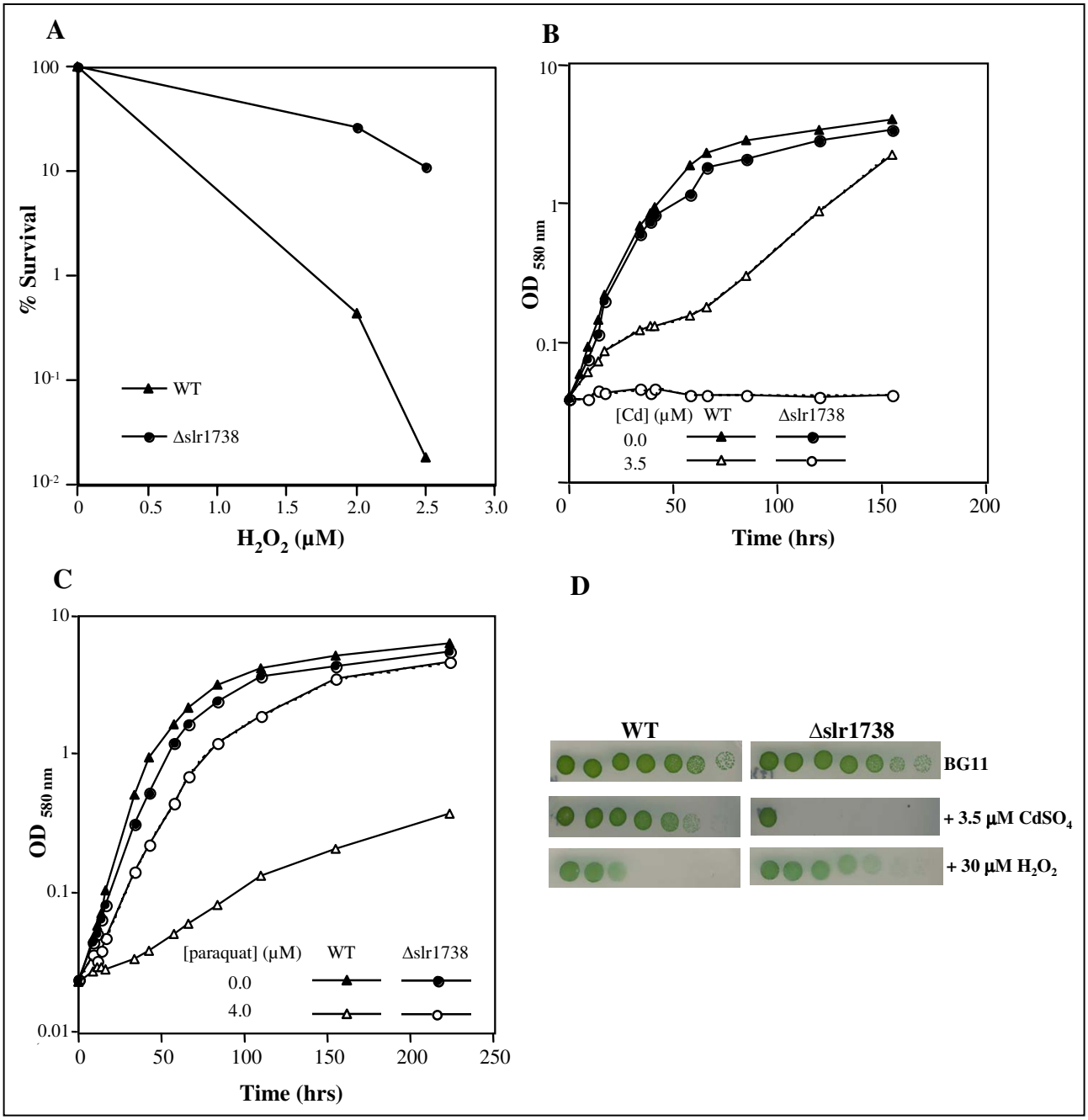


Figure 5

Influence of the slr1738 gene on the tolerance to cadmium and oxidative stress. Panel A: Typical survival to H_2O_2 of the wild type (WT, triangles) and Δslr1738 mutant (circles) cells. Panels B, C and D: typical growth of the WT and Δslr1738 cells cultivated in liquid (Panels B and C) or solid (Panel D) BG11 media with or without H_2O_2, paraquat and $CdSO_4$ at the indicated concentration. These experiments were repeated three times.

relevant assays (see Figs. 1 to 5) performed with appropriate metal doses that varied with the number of cells to be treated, and/or attested by the following evidences. First, in every case all gene members of the same operon were found to be co-regulated. Second, all Fe- and Zn-acquisition genes responded the expected way to the availability of their cognate metal. Third, most of the dispersed genes encoding the protein-subunits of the same complex were

found to be co-regulated, as observed for instance in the case of the photosynthesis (PS) genes (see Additional file 4 panel B). Fourth, in agreement with the Cd-mediated downregulation of most genes PS and the concomitant upregulation of protein-degradation genes (see Additional file 4 panel A), we showed that Cd decreases both the abundance (Fig. 3) and activity (See oxygen evolution data in the Results section) of the PS machinery. Fifth, consistent with the Cd-elicited induction of the arsenate reductase gene *arsC* (see Additional file 4 panel C) we showed that *arsC* operates in Cd tolerance (Fig. 4). Thus, the ArsC enzyme has a great biotechnological potential in contributing to the tolerance to two widespread persistent pollutants: arsenic and cadmium. Possibly the ArsC enzyme, which employed glutathione (GSH) and glutaredoxin as reductants [45], could somehow "sequester" Cd in generating an hypothetical GSH-Cd complex less toxic than Cd. Sixth, as inferred from the Cd-elicited upregulation of numerous high light-inducible genes (see Additional file 4 panel A) we showed that Cd decreases the tolerance to light (Fig. 2C). Seventh, as anticipated from its Cd-elicited induction (Fig. 1 and see Additional file 4 panel C) we demonstrated that the Slr1738 regulator mediates several responses that are crucial to protection against Cd, such as the decline of the PS machinery and the upregulation of the *arsC* gene (Fig. 3, Fig. 4 and Fig. 5 and see Additional file 4).

The occurrence of large clusters of co-regulated genes, encoding ribosome, ATPase or Fe uptake proteins, suggests a mechanism of global control of gene expression involving chromosomal structure, similarly to chromatin remodeling in eukaryotic cells. This prediction is comforted by the findings (see Additional file 2) that the *Synechocystis* HU and Dps nucleoprotein genes, possibly involved in such structure-dependent global regulation [54], are regulated by Cd (see Additional file 4), positively (HU, sll1712) or negatively (Dps, slr1894).

Many of our data support the notion of metal selectivity. For examples, (i) Fe but not Zn mimicked the Cd-mediated downregulation of N acquisition and assimilation genes (see Additional file 4 panel F); and (ii) Zn but not Fe mimicked the Cd-mediated decline of the PS machinery (Fig. 3 and see Additional file 4) and the downregulation of ribosomal genes which is therefore not a general stress response. By contrast, numerous genes responded the same way to Cd, Fe, Zn and H_2O_2 (see Additional file 4), suggesting that reactive oxygen species might act in signal transduction of stress responses, as proposed [55].

Both H_2O_2 and Cd were found to upregulate numerous genes operating in tolerance to oxidative stress (see Additional file 4 panel D) and to the related high light stress (see Additional file 4 panel B) that also generates toxic reactive oxygen species (ROS) [23]. These findings, which were anticipated in the case of the H_2O_2 stress, were confirmed by showing that both H_2O_2 and Cd render cells light sensitive (Fig. 2C). In addition, both H_2O_2 and Cd upregulated the *suf* genes (see Additional file 4 panel D) involved in iron-sulfur cluster biogenesis or repair [39] and the Fe uptake genes (see Additional file 4 panel C; all genes in the case of H_2O_2 and half of them in the case of Cd). Consistently, we found that increasing the concentration of Fe in the medium increases the cell tolerance to both H_2O_2 and Cd (Fig. 2). These results are reminiscent to what occurs in oxidative-stressed *E. coli* cells [40-42]. They suggest that *Synechocystis* challenged with either H_2O_2 or Cd accelerates Fe uptake strongly (H_2O_2) or moderately (Cd) to provide extra Fe atoms for the repair of damaged Fe-S clusters. Having also noticed that Cd triggers a larger breakdown of the Fe-rich PS machinery than H_2O_2 (Fig. 3), we believe that cells challenged with H_2O_2 or Cd use two strategies to provide extra Fe atoms to the machinery that synthesizes or repairs Fe-S centers. H_2O_2-treated cells undergoing a limited PS-decline (Fig. 3D) mostly accelerate the intake of Fe from the medium, while Cd-stressed cells that moderately increase Fe intake breakdown a part of their abundant PS machinery (Fig. 3A), which normally contains 21–23 Fe atoms per PS unit [43]. Consequently, we predicted, and confirmed, that increasing the availability of Fe limits the Cd-elicited decline of the PS-machinery (Fig. 3).

Many of the key genes involved in acquisition and metabolism of C, N and S that were downregulated by Cd and H_2O_2 are coding for ATP-consuming enzymes (see Additional file 4 panels E and F). These responses can be viewed as an ATP-sparing process used by the cells to compensate for the decreased production of ATP caused by the decline of the PS apparatus (Cd and to a lesser extent H_2O_2) or of the respiration machinery (H_2O_2 but not Cd downregulates cytochrome oxidase genes, See TableS4B). Similarly, the downregulation of ribosomal genes (see Additional file 4 panel A) encoding normally abundant proteins [22] triggered by cells facing Cd and H_2O_2 is likely aimed at sparing C, N and S nutrients to compensate for the downregulation of the corresponding acquisition and assimilation genes.

Based on the presently reported findings, we view the responses to a continuous Cd stress as a two temporal-phases process. In the early phase occurring during the first 60 min of exposure (Table 1), Cd-stressed cells regulate mainly the genes operating in metal transport (see Additional file 4 panel C) and protein maturation and degradation (see Additional file 4 panel A). These responses presumably limit Cd entry into the cells and incorporation in place of the cognate metal cofactor of metalloproteins. In prolonged exposures (after 60 min.),

these regulations are conserved, and even amplified in term of the number of responsive processes, thereby defining the next phase designated as "massive" for this reason. At this stage, the responses to Cd can be viewed as an integrated "Yin Yang" reprogramming of the whole cellular metabolism. As the Yin process, most key genes operating in uptake and assimilation of inorganic nutrients (C, N and S) and protein synthesis are turned down. These responses allow (i) the sparing of both energy (ATP) and reducing power (NAD(P)H) normally consumed by nutrient assimilation and subsequent metabolism, and (ii) the limitation of the poisoning incorporation of Cd in metalloenzymes. As the compensatory Yang process, the PS breakage of the PS machinery, which decreases the production of both ATP and NADPH, liberates Fe and C, N and S nutrient assimilates that can be recycled into the synthesis of Cd-tolerance enzymes such as the ArsC arsenate reductase (Fig. 4) and, presumably, other Cd-inducible enzymes: Suf proteins (see above), flavodoxin (IsiB), ferredoxin (FedII), flavoproteins (Flv2 and Flv4), glutathione peroxidase (Gpx1), peroxiredoxin (Ahpc-like), thioredoxin (TrxA), hydrogenase subunits (HypA, D, E) and the ZiaA (the ATPase homologous to the cadmium export ATPase of other organisms). Furthermore, other new Cd-tolerance enzymes might be discovered in the future, among the product of the orphan genes which apperaed to be upregulated by Cd (see Additional files 1 and 3). We showed that the Cd-induced Slr1738 regulator (Fig. 1 and TableS4) plays a central role in the protection against Cd (Fig. 5) in mediating several of the important regulations, such as the breakage of the PS machinery, the downregulation of ribosomal genes, and the upregulation of the *arsC* arsenate reductase and *suf* genes.

Conclusion

Using the cyanobacterium *Synechocystis* PCC6803 as a model organism, we analyzed the global responses of environmentally important cells to stresses triggered by Cd (an abundant persistant pollutant), H_2O_2 (the paradigm ROS agent), or drastic changes in Fe availability, which appeared to modulate the tolerance to Cd and H_2O_2. Our results indicate that cells challenged with H_2O_2 or Cd use different strategies for the same purpose of increasing the supply of Fe atoms to the synthesis and repair of Fe-requiring metalloenzymes. While H_2O_2-challenged cells preferentially accelerate Fe intake, Cd-stressed cells preferentially breakdown the Fe-rich PS machinery to liberate Fe atoms. We view the responses to Cd as an integrated "Yin Yang" metabolic reprogramming. As the "Yin" process, the ATP- and nutrients-sparing downregulation of anabolism limits the poisoning incorporation of Cd into metalloenzymes. As the compensatory "Yang" process, the PS breakdown liberates nutrient assimilates for the synthesis of Cd-tolerance proteins. We found that this reprogramming is mediated by the Slr1738 transcriptional regulator that also operates in oxidative stress tolerance, in agreement with its sequence homology with the peroxide resistance regulator of *Bacillus subtilis*. Further studies will be necessary to understand the influence of Cd and H_2O_2 on the activity of Slr1738.

Methods

Bacterial strains, growth and survival analyses

The unicellular cyanobacterium *Synechocystis* PCC6803 was grown as described [56] at 30°C in liquid BG11 medium enriched with Na_2CO_3 (3.78 mM, final concentration), under continuous white light of standard fluence (2,500 luxes, i.e. 31.25 $\mu E.m^{-2}.s^{-1}$). When required kanamycin 50–300 $\mu g.ml^{-1}$ was added to the cultures. For growth assays, cells grown three times up to mid log phase (OD_{580} 0.5 units, i.e. 2.5×10^7 $cells.ml^{-1}$) were inoculated into fresh BG11 medium with or without $CdSO_4$ or paraquat (at the indicated concentration), and OD_{580} were measured at time intervals. For survival analysis, cells in mid log phase were incubated for 1 h with H_2O_2 (at the indicated concentration), washed twice with BG11, plated on solid BG11, and the colonies were counted after 5–7 days of incubation under standard conditions. The influence of various agents on the growth and survival on solid media was assayed as follows. Four fold serial dilutions of liquid cultures were spotted as 15 µl dots onto BG11 plates with or without the indicated concentration of the tested agents. Plates were then incubated for 4–5 days under standard conditions prior to scanning. For survival analysis, cells were harvested, washed and resuspended in BG11 medium prior to plating and counting.

Construction of knockout mutants of Slr0946 (arsenate reductase) and Slr1738 regulator

Specific oligonucleotides were used for PCR amplification from the *Synechocystis* genome [57] of each studied gene along with its two 0.3 kb-long flanking DNA segments that serve as platforms for homologous recombinations mediating targeted gene replacement [58]. The PCR products were independently cloned in the pGEMt plasmid (Pharmacia), and inactivated as follows. For slr0946, an internal 223 bp segment (starting from the 7th nucleotide downstream of its GTG start codon) was substituted by the *Stu*I restriction site that was subsequently used to insert the *Hinc*II Km^r cassette originating from the pUC4K plasmid (Pharmacia). For slr1738 inactivation, the 388 bp segment beginning 6 nucleotides behind the ATG start codon was replaced by the *Sma*I site in which we cloned the *Hinc*II Km^r cassette. The resulting deletion cassettes slr0946::Km^r and slr1738::Km^r were sequenced (Big Dye kit, ABI Perking Elmer) prior to transformation to *Synechocystis*. Through PCR and sequence analyses, we verified that the antibiotic resistant marker had been inserted properly in the genome of the transformant clones, i.e. in

place of the corresponding studied gene. These nullmutants grew healthy in standard conditions, demonstrating that both Slr0946 and Slr1738 proteins are dispensable to the viability of *Synechocystis*.

Photosynthetic pigments determination by absorption spectrometry

Absorption spectra of whole cells grown or challenged on solid medium and resuspended in water, were monitored with a DU640 spectrophotometer (Beckman). Samples were adjusted for equal scattering at 800 nm. Carotenoids absorb light between 350 and 540 nm. The absorption maximum for phycocyanin is 630 nm and that for chlorophyll a are 442 nm and 681 nm. These experiments were repeated at least three times.

RNA isolation

Because of their short half lives typical of prokaryotic transcripts, *Synechocystis* mRNA were rapidly prepared (in less than 2 min.) from cells grown or challenged on solid media as described [56]. Briefly, 300 ml of mid-log phase liquid cultures (2.5×10^7cells.ml^{-1}) grown in standard conditions were rapidly concentrated 40-fold by centrifugation and spotted as 20 µl dots on BG11 plates with, or without (control samples), $CdSO_4$ (50 µM); $(NH_4)FeH_2C_6H_5O_7$ (17 µM); $ZnSO_4$ (776 µM); or H_2O_2 (3 mM). Then, plates were incubated for the indicated times (in min) prior to cell harvest and fast disruption with an Eaton press. Iron depletion analyses were performed with liquid cell suspensions to avoid uncontrolled liberation of Fe atoms from agar. Hence, exponentially growing cells washed in Fe-free medium were challenged for 48 h in liquid medium containing 0 to 2 µM of $(NH_4)FeH_2C_6H_5O_7$. Then, cells were harvested by centrifugation and resuspension and disrupted. RNA were extracted [56] with the RNeasy kit from Qiagen (DNA microarrays kit) and treated with RNase-free DNase I (Roche). The RNA concentration and purity were determined by A260 and A280 measurement (A260/A280 > 1.9), as well as by migration on agarose gel to verify the absence of RNA degradation.

DNA-microarray data acquisition and statistical analysis

The microarrays data presently reported have been deposited in the MIAME compliant NCBIs Gene Expression Omnibus [59] under the accession number GSE3755 (see Additional file 6) DNA microarrays (IntelliGene™ CyanoCHIP version 1.2 or 2.0, Takara), covering 2,891 (CyanoCHIP1.2) or 2,954 (CyanoCHIP2.0) of the 3,168 ORFs of *Synechocystis* were purchased from Cambrex Bio Science and manipulated as described [56]. Test RNA (from stressed cells) and corresponding control RNA (untreated cultures) were reverse transcribed, differentially labeled with Cy3 and Cy5 dyes and hybridized in a replicate dye swap. Arrays were immediately scanned with a GenePix™ 4000B scanner (Axon Instruments), and images were analyzed with GenePixTM Pro 4.0 (Molecular Devices). Spots were considered when they lack blemishes, deformations or dusts, and their fluorescence signal exceeded the local background plus 2 standard deviations. Then, signal intensity was determined by subtracting local background of each spot (GenePix™ Pro 4.0). Each GenePix Result file (.GPR) was converted to a TIGR Array viewer file (.TAV) using TIGR ExpressConverter version 1.7 for signal analysis. All spot intensities have been normalized with the LOWESS method [60], using the locfit function of the TIGR Midas version 2.19 [61] with the smooth parameter set to 0.33 as recommended [62]. Normalized measures served to compute the ratios of Cy3/Cy5 intensity and the associated log2-transform (denote log2-ratios) for each gene. For each replicated dye-swap, the average expression ratio of a given gene is calculated as the geometric mean of the two ratios [62].

Three lines of evidences attested the quality of our data. First our normalization method was validated with both internal (positive: *Synechocystis* DNA; negative: Salmon sperm DNA) and external (human TFR mRNA) controls. Second, for each dye-swap, correlation coefficients calculated between both replicates (see Additional file 5) appeared to be greater than 0.9 in most cases, thereby attesting the within-study reproducibility. Third, to analyze the within-platform variations, Cy3- and Cy5-labeled cDNAs were prepared from a single preparation of total RNA from unstressed cells, mixed together and hybridized to a microarray. The distribution of expression ratio ((see Additional files 4 and 6) showed that 90% of them fall within the range (0.80 – 1.25) and 99% within the range (0.64 – 1.57). Moreover, the mean of this distribution is equal to 1.00 (standard deviation equal to 0.13). Consequently, we felt confident to regard as regulated any particular gene the expression level of which was changed at least 1.9 fold.

Identification of the two temporal phases of the responses to Cd and H_2O_2

We have considered each cDNA array as a split replicate, without averaging the dye-swap values. For each stress, the microarray data were dispatched in two groups each corresponding to a presumed kinetic phase. In the case of Cd the first group of data (15 mins to 60 mins) contains 8 replicates and the second (90 min to 360 min) contains 10 replicates, while for H_2O_2 the first group (15 min to 30 min) contains 4 replicates and the second (180 min to 420 min) contains 6 replicates. To identify genes with log2-ratios significantly different between the two time phases, p-values were first calculated for each gene using a moderated t-test based on an empirical Bayes analysis that is equivalent to shrinkage (or expansion) of the estimated sample variances towards a pooled estimate, resulting in a

more stable inference. The p-values of the t-test were adjusted for multiple hypotheses testing, controlling the false discovery rate (FDR) as proposed by [63]. Thus, using a cut-off of the adjusted p-values at 0.05 gives and approximate level of False Discovery Rate (FDR) at 0.05. Using a strict cut-off of p = 1e-3 we found 791 genes differentially expressed in the two kinetic phases of responses to Cd (see Additional file 2) and 228 phase-responsive genes in the case of H_2O_2 (TableS3). The statistical analysis was carried out in the R language release 2.2, using the package limma [64] from the Bioconductor project [65].

Measurement of photosynthetic activity

Cells incubated for 3 and 6 h on solid BG11 medium with or without CdSO4 (50 μM) were washed and resuspended in BG11 medium as described in the RNA isolation section. Photosynthetic oxygen-evolving activity of intact cells was measured at 30°C under saturating light intensity with a Clark-type oxygen electrode (Hansatech).

Over-expression and purification of the Slr1738 protein fused to a hexahistidine tag

The Slr1738 coding sequence was PCR amplified from the *Synechocystis* genome, using appropriate oligonucleotide primers to embed its ATG initiation codon into a *Nde* I restriction site and introduce a *Bam* HI site behind its stop codon. The resulting *Nde* I-*Bam* HI restriction fragment was cloned into the pET28 (+) *E. coli* expression vector opened with the same enzymes, thereby allowing the in-frame fusion of the 6 × His tag with the Slr1738 amino acids sequence. After sequence verification (Big Dye kit, ABI Perking Elmer) the pET28-1738 plasmid was transformed into *E. coli* BL21 (DE3) selecting for resistance to kanamycin (50 $\mu g.ml^{-1}$). Transformant cells were grown at 37°C in Km-containing Luria Bertani medium up to an optical density (A600) of 0.8. At that time, 1 mM isopropyl-thio-β-D-galactopyranoside (IPTG) was added to induce the synthesis of the 6 × His-Slr1738 protein, and cells were further incubated for 15 h at 30°C, harvested by centrifugation and resuspended in 20 ml of 20 mM Tris pH 8.0, 500 mM NaCl and 5 mM imidazole (lysis buffer). Cells were disrupted by sonication (Microson), centrifuged at 14,000 g for 20 min at 4°C, and the supernatant was applied to a nickel-nitrilotriacetic acid-agarose column (3 ml) equilibrated with 25 ml of lysis buffer. After washings with 30 ml of lysis buffer and buffer A (20 mM Tris pH 8.0, 500 mM NaCl and 50 mM imidazole), recombinant proteins were eluted with 6 ml of buffer B (20 mM Tris pH 8.0, 500 mM NaCl and 500 mM imidazole). 6His-Slr1738 containing fractions were pooled, desalted on a PD10 Sephadex G-25M column (Amersham Biosciences). The Purity of the 6His-Slr1738 protein was greater than 95%, as judged by SDS-PAGE electrophoresis.

Western blot analysis of selected proteins

Crude cell extract (5 μg) of *Synechocystis* cells incubated on solid media with or without $CdSO_4$ (50 μM, 360 min.) or H_2O_2 (3 mM, 30 min.) were harvested, disrupted (see above), electrophoresed on 13% SDS-PAGE [66] and transferred onto nitrocellulose sheets as described [67]. For detections we use the following rabbit antibodies: anti-Slr1738 (this work, dilution 1:20000); anti-psaC (dilution 1:1000) or anti-rbcL (dilution 1:5000) from Agrisera; anti-IsiA (kindly provided by Dr. A. Wilde, dilution 1:5000); anti-IsiB (kindly provided by Dr. M. Hagemann, dilution 1:5000). Horseradish peroxidase-conjugated goat anti-rabbit antibodies (dilution 1:4000) were used as second antibody, and immune complexes were revealed by chemiluminescence (ECL kit, Amersham Biosciences).

Authors' contributions

LH participated in the design and realization of the transcriptome experiments in the wild type strain. MF constructed the slr1738 null mutant and carried out the transcriptome and phentotypic analysis of this strain. BM carried out the construction and phenotypic analyses of the *ars*C-null mutant, and participated to other cell-fitness analyses. MM participated in the bioinformatics analysis of the transcriptome data, and in the drafting of the relevant part of the manuscript. AP carried out the overproduction of Slr1738 protein and corresponding antibodies and performed the Western blot experiments. PL participated in the analysis and interpretation of the transcriptome data and in the drafting of the whole manuscript. JCA participated in the bioinformatics analysis of the transcriptome data, and in the drafting of the relevant part of the manuscript. CCC participated in the conception, acquisition, analysis and interpretation of all data; carried out the oxygen evolution measurements; and participated in the drafting of the whole manuscript. FC participated in conception and supervision of the whole study, and wrotes the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Analysis of the unstressed versus unstressed control experiment. These file describe the distribution of allexpression ratio for the unstressed versus unstressed control experiment, showing the extreme, mean and quantile values.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-8-350-S1.xls]

Additional file 2

Kinetic analysis of the transcriptional responses to cadmium. The data provided indicate for each gene (column 1), encoding the indicated protein (column 12), the log-ratio of response to the indicated duration of the cadmium treatment (columns 2 to 10) and the p-value obtained by the analysis (column 11).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-350-S2.xls]

Additional file 3

Kinetic analysis of the transcriptional responses to hydrogen peroxide. The data provided indicate for each gene (column 1), encoding the indicated protein (column 8), the log-ratio of response to the indicated duration of the hydrogen peroxide treatment (columns 2 to 6), and the p-value obtained by the analysis (column 7).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-350-S3.xls]

Additional file 4

Relevant List of Stress-Responsive Genes. The data provided indicate for each gene (extreme left and right columns), sorted by the physiological function they operate in, the log-ratio of response to the indicated duration of stresses.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-350-S4.xls]

Additional file 5

Dye-Swap-correlations. The data provided indicate (for each time of the 5 kinetics) the correlation coefficient between the two ratio samples obtained with the two microarrays of each dye-swap.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-350-S5.xls]

Additional file 6

GSM numbers of all microarray experiments. This table provides the links for accessing our microarray data which have been deposited in the GEO website.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-350-S6.xls]

Acknowledgements

This work was supported by the French scientific Programs "Toxicologie Nucléaire Environnementale" and "ANR Biosys06_134823: SULFIRHOM". L.H, B.M, M.M and A.P were recipients of fellowships from the CEA (France). M.F. was recipient of MENESR PhD fellowship. We thank A. Wilde and M. Hagemann for their kind gift of antibodies directed against the IsiA and IsiB proteins, respectively; and C. Creminon and J-C. Robillard for their help in the preparation of antibodies directed against Slr1738.

References

1. Partensky F, Hess WR, Vaulot D: **Prochlorococcus, a marine photosynthetic prokaryote of global significance.** *Microbiol Mol Biol Rev* 1999, **63:**106-127.
2. Zehr JP, Waterbury JB, Turner PJ, Montoya JP, Omoregie E, Steward GF, Hansen A, Karl DM: **Unicellular cyanobacteria fix N2 in the subtropical North Pacific Ocean.** *Nature* 2001, **412:**635-638.
3. Nishiyama Y, Yamamoto H, Allakhverdiev SI, Inaba M, Yokota A, Murata N: **Oxidative stress inhibits the repair of photodamage to the photosynthetic machinery.** *EMBO J* 2001, **20(20):**5587-5594.
4. Satarug S, Baker JR, Urbenjapol S, Haswell-Elkins M, Reilly PE, Williams DJ, Moore MR: **A global perspective on cadmium pollution and toxicity in non-occupationally exposed population.** *Toxicol Lett* 2003, **137(1-2):**65-83.
5. Andrew AS, Warren AJ, Barchowsky A, Temple KA, Klei L, Soucy NV, O'Hara KA, Hamilton JW: **Genomic and proteomic profiling of responses to toxic metals in human lung cells.** *Environ Health Perspect* 2003, **111(6):**825-835.
6. Waisberg M, Joseph P, Hale B, Beyersmann D: **Molecular and cellular mechanisms of cadmium carcinogenesis.** *Toxicology* 2003, **192(2-3):**95-117.
7. Rosenzweig AC: **Metallochaperones: bind and deliver.** *Chem Biol* 2002, **9(6):**673-677.
8. Bryant DA: **The molecular biology of cyanobacteria.** In *Advances in photosynthesis Volume 1*. Edited by: Govindjee . Dordrecht , Kluwer academic publishers; 1994.
9. Stohs SJ, Bagchi D: **Oxidative mechanisms in the toxicity of metal ions.** *Free Radic Biol Med* 1995, **18(2):**321-336.
10. Ferris MJ, Palenik B: **Niche adaptation in ocean cyanobacteria.** *Nature* 1998, **396:**226-228.
11. Peschek GA: **Structure-function relationships in the dual-function photosynthetic-respiratory electron-transport assembly of cyanobacteria (blue-green algae).** *Biochem Soc Trans* 1996, **24(3):**729-733.
12. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D: **Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus.** *Proc Natl Acad Sci U S A* 2002, **99:**12246-12251.
13. Gray MW: **Origin and evolution of organelle genomes.** *Curr Opin Genet Dev* 1993, **3(6):**884-890.
14. Drazkiewicz M, Tukendorf A, Baszynski T: **Age-dependent response of maize leaf segments to cadmium treatment: effect on chlorophyll fluorescence and phytochelatin accumulation.** *J Plant Physiol* 2003, **160(3):**247-254.
15. Carrier P, Baryla A, Havaux M: **Cadmium distribution and microlocalization in oilseed rape (Brassica napus) after long-term growth on cadmium-contaminated soil.** *Planta* 2003, **216(6):**939-950.
16. Bachmann T: **Transforming cyanobacteria into bioreporters of biological relevance.** *Trends Biotechnol* 2003, **21(6):**247-249.
17. Gong R, Ding Y, Liu H, Chen Q, Liu Z: **Lead biosorption and desorption by intact and pretreated spirulina maxima biomass.** *Chemosphere* 2005, **58(1):**125-130.
18. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S: **Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions.** *DNA Res* 1996, **3:**109-136.
19. Mazouni K, Bulteau S, Cassier-Chauvat C, Chauvat F: **Promoter element spacing controls basal expression and light-inducibility of the cyanobacterial secA gene.** *Mol Microbiol* 1998, **30:**1113-1122.
20. Poncelet M, Cassier-Chauvat C, Leschelle X, Bottin H, Chauvat F: **Targeted deletion and mutational analysis of the essential (2Fe-2S) plant-like ferredoxin in Synechocystis PCC6803 by plasmid shuffling.** *Mol Microbiol* 1998, **28:**813-821.
21. Mazouni K, Domain F, Cassier-Chauvat C, Chauvat F: **Molecular analysis of the key cytokinetic components of cyanobacteria: FtsZ, ZipN and MinCDE.** *Mol Microbiol* 2004, **52(4):**1145-1158.
22. Mrazek J, Bhaya D, Grossman AR, Karlin S: **Highly expressed and alien genes of the Synechocystis genome.** *Nucleic Acids Res* 2001, **29(7):**1590-1601.
23. Munekage Y, Hashimoto M, Miyake C, Tomizawa K, Endo T, Tasaka M, Shikanai T: **Cyclic electron flow around photosystem I is essential for photosynthesis.** *Nature* 2004, **429(6991):**579-582.

24. Cavet JS, Borrelly GP, Robinson NJ: **Zn, Cu and Co in cyanobacteria: selective control of metal availability.** *FEMS Microbiol Rev* 2003, **27(2-3):**165-181.
25. van Waasbergen LG, Dolganov N, Grossman AR: **nblS, a gene involved in controlling photosynthesis-related gene expression during high light and nutrient stress in Synechococcus elongatus PCC 7942.** *J Bacteriol* 2002, **184(9):**2481-2490.
26. Hihara Y, Kamei A, Kanehisa M, Kaplan A, Ikeuchi M: **DNA microarray analysis of cyanobacterial gene expression during acclimation to high light.** *Plant Cell* 2001, **13(4):**793-806.
27. Huang L, McCluskey MP, Ni H, LaRossa RA: **Global Gene Expression Profiles of the Cyanobacterium Synechocystis sp. Strain PCC 6803 in Response to Irradiation with UV-B and White Light.** *J Bacteriol* 2002, **184(24):**6845-6858.
28. Tu CJ, Shrager J, Burnap RL, Postier BL, Grossman AR: **Consequences of a deletion in dspA on transcript accumulation in Synechocystis sp. strain PCC6803.** *J Bacteriol* 2004, **186(12):**3889-3902.
29. He Q, Dolganov N, Bjorkman O, Grossman AR: **The high light-inducible polypeptides in Synechocystis PCC6803. Expression and function in high light.** *J Biol Chem* 2001, **276(1):**306-314.
30. Yeremenko N, Kouril R, Ihalainen JA, D'Haene S, van Oosterwijk N, Andrizhiyevskaya EG, Keegstra W, Dekker HL, Hagemann M, Boekema EJ, Matthijs HC, Dekker JP: **Supramolecular organization and dual function of the IsiA chlorophyll-binding protein in cyanobacteria.** *Biochemistry* 2004, **43(32):**10308-10313.
31. Silva P, Thompson E, Bailey S, Kruse O, Mullineaux CW, Robinson C, Mann NH, Nixon PJ: **FtsH is involved in the early stages of repair of photosystem II in Synechocystis sp PCC 6803.** *Plant Cell* 2003, **15(9):**2152-2164.
32. Schneider D, Berry S, Volkmer T, Seidler A, Rogner M: **PetC1 is the major Rieske iron-sulfur protein in the cytochrome b6f complex of Synechocystis sp. PCC 6803.** *J Biol Chem* 2004, **279(38):**39383-39388.
33. Thelwell C, Robinson NJ, Turner-Cavet JS: **An SmtB-like repressor from Synechocystis PCC 6803 regulates a zinc exporter.** *Proc Natl Acad Sci U S A* 1998, **95(18):**10728-10733.
34. Garcia-Dominguez M, Lopez-Maury L, Florencio FJ, Reyes JC: **A gene cluster involved in metal homeostasis in the cyanobacterium Synechocystis sp. strain PCC 6803.** *J Bacteriol* 2000, **182(6):**1507-1514.
35. Rensing C, Ghosh M, Rosen BP: **Families of soft-metal-ion-transporting ATPases.** *J Bacteriol* 1999, **181(19):**5891-5897.
36. Raux E, Lanois A, Warren MJ, Rambach A, Thermes C: **Cobalamin (vitamin B12) biosynthesis: identification and characterization of a Bacillus megaterium cobI operon.** *Biochem J* 1998, **335 (Pt 1):**159-166.
37. Katoh H, Hagino N, Grossman AR, Ogawa T: **Genes essential to iron transport in the cyanobacterium Synechocystis sp. strain PCC 6803.** *J Bacteriol* 2001, **183(9):**2779-2784.
38. Singh AK, McIntyre LM, Sherman LA: **Microarray analysis of the genome-wide response to iron deficiency and iron reconstitution in the cyanobacterium Synechocystis sp. PCC 6803.** *Plant Physiol* 2003, **132(4):**1825-1839.
39. Wang T, Shen G, Balasubramanian R, McIntosh L, Bryant DA, Golbeck JH: **The sufR gene (sll0088 in Synechocystis sp. strain PCC 6803) functions as a repressor of the sufBCDS operon in iron-sulfur cluster biogenesis in cyanobacteria.** *J Bacteriol* 2004, **186(4):**956-967.
40. Benov L, Fridovich I: **Growth in iron-enriched medium partially compensates Escherichia coli for the lack of manganese and iron superoxide dismutase.** *J Biol Chem* 1998, **273(17):**10313-10316.
41. Zheng M, Wang X, Templeton LJ, Smulski DR, LaRossa RA, Storz G: **DNA microarray-mediated transcriptional profiling of the Escherichia coli response to hydrogen peroxide.** *J Bacteriol* 2001, **183(15):**4562-4570.
42. Djaman O, Outten FW, Imlay JA: **Repair of oxidized iron-sulfur clusters in Escherichia coli.** *J Biol Chem* 2004, **279(43):**44590-44599.
43. Straus NA: **Iron deprivation: Physiology and Gene Regulation.** In *The Molecular Biology of Cyanobacteria* Edited by: Bryant DA. Dordrecht , Kluwer Academic Publisher; 1994:731-750.
44. Lopez-Maury L, Florencio FJ, Reyes JC: **Arsenic sensing and resistance system in the cyanobacterium Synechocystis sp. strain PCC 6803.** *J Bacteriol* 2003, **185(18):**5363-5371.
45. Li R, Haile JD, Kennelly PJ: **An arsenate reductase from Synechocystis sp. strain PCC 6803 exhibits a novel combination of catalytic characteristics.** *J Bacteriol* 2003, **185(23):**6780-6789.
46. Badger MR, Price GD: **CO2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution.** *J Exp Bot* 2003, **54(383):**609-622.
47. Figge RM, Cassier-Chauvat C, Chauvat F, Cerff R: **Characterization and analysis of an NAD(P)H dehydrogenase transcriptional regulator critical for the survival of cyanobacteria facing inorganic carbon starvation and osmotic stress.** *Mol Microbiol* 2001, **39:**455-469.
48. Aboulmagd E, Oppermann-Sanio FB, Steinbuchel A: **Purification of Synechocystis sp. strain PCC6308 cyanophycin synthetase and its characterization with respect to substrate and primer specificity.** *Appl Environ Microbiol* 2001, **67(5):**2176-2182.
49. Garcia-Dominguez M, Reyes JC, Florencio FJ: **NtcA represses transcription of gifA and gifB, genes that encode inhibitors of glutamine synthetase type I from Synechocystis sp. PCC 6803.** *Mol Microbiol* 2000, **35(5):**1192-1201.
50. Kobayashi M, Ishizuka T, Katayama M, Kanehisa M, Bhattacharyya-Pakrasi M, Pakrasi HB, Ikeuchi M: **Response to oxidative stress involves a novel peroxiredoxin gene in the unicellular cyanobacterium Synechocystis sp. PCC 6803.** *Plant Cell Physiol* 2004, **45(3):**290-299.
51. Li H, Singh AK, McIntyre LM, Sherman LA: **Differential gene expression in response to hydrogen peroxide and the putative PerR regulon of Synechocystis sp. strain PCC 6803.** *J Bacteriol* 2004, **186(11):**3331-3345.
52. Hosoya-Matsuda N, Motohashi K, Yoshimura H, Nozaki A, Inoue K, Ohmori M, Hisabori T: **Anti-oxidative stress system in cyanobacteria. Significance of type II peroxiredoxin and the role of 1-Cys peroxiredoxin in Synechocystis sp. strain PCC 6803.** *J Biol Chem* 2005, **280(1):**840-846.
53. Bsat N, Herbig A, Casillas-Martinez L, Setlow P, Helmann JD: **Bacillus subtilis contains multiple Fur homologues: identification of the iron uptake (Fur) and peroxide regulon (PerR) repressors.** *Mol Microbiol* 1998, **29(1):**189-198.
54. Dorman CJ, Deighan P: **Regulation of gene expression by histone-like proteins in bacteria.** *Curr Opin Genet Dev* 2003, **13(2):**179-184.
55. Apel K, Hirt H: **Reactive oxygen species: metabolism, oxidative stress, and signal transduction.** *Annu Rev Plant Biol* 2004, **55:**373-399.
56. Domain F, Houot L, Chauvat F, Cassier-Chauvat C: **Function and regulation of the cyanobacterial genes lexA, recA and ruvB: LexA is critical to the survival of cells facing inorganic carbon starvation.** *Mol Microbiol* 2004, **53(1):**65-80.
57. Nakamura Y, Kaneko T, Hirosawa M, Miyajima N, Tabata S: **CyanoBase, a www database containing the complete nucleotide sequence of the genome of Synechocystis sp. strain PCC6803.** *Nucl Acids Res* 1998, **26:**63-67.
58. Labarre J, Chauvat F, Thuriaux P: **Insertional mutagenesis by random cloning of antibiotic resistance genes into the genome of the cyanobacterium Synechocystis PCC6803.** *J Bacteriol* 1989, **171:**3449-3457.
59. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30(1):**207-210.
60. Cleveland WS: **Robust locally weighted regression and smoothing scatterplots.** *J Amer Stat Assoc* 1979, **74:**829-836.
61. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J: **TM4: a free, open-source system for microarray data management and analysis.** *Biotechniques* 2003, **34(2):**374-378.
62. Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet* 2002, **32 Suppl:**496-501.
63. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society* 1995, **57:**289-300.
64. Gordon KS: **Limma: linear models for microarray data.** In *Bioinformatics and Computational Biology Solutions using R and Bioconductor* Edited by: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New York , Springer; 2005:397-420.

65. Gentleman R, Carey VJ, Bates DM, Bolstad BM, Dettlings M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **580:** [http://genomebiology.com/2004/5/10/R80].
66. Chua NH: **Electrophoresis analysis of chloroplast proteins.** *Volume 69*. Academic Press, INC.; 1980:434-436.
67. Towbin H, Staehelin T, Gordon J: **Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications.** *Proc Natl Acad Sci U S A* 1979, **76(9):**4350-4354.

Mining Biological Data using Pyramids

Géraldine Polaillon[1], Laure Vescovo[1], Magali Michaut[2], and Jean-Christophe Aude[2]

[1] Département Informatique, Supélec
Plateau de Moulon, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette cedex, France, *geraldine.polaillon@supelec.fr, laure.vescovo@supelec.fr*

[2] Service de Biologie Intégrative et de Génétique Moléculaire, CEA
CEA Saclay, 91191 Gif-sur-Yvette cedex, France, *magali.michaut@cea.fr, jean-christophe.aude@cea.fr*

Abstract. This paper is a review of promising applications of pyramidal classification to biological data. We show that overlapping and ordering properties can give new insights that can not be achieved using more classical methods. We examplify our point using three applications: (i) a genome scale sequence analysis, (ii) a new progressive multiple sequence alignment method, (iii) a cluster analysis of transcriptomic data.

1 Introduction

Biology has always benefited from advances in mathematics, more specifically in statistics and classification. Conversely, mathematical discoveries are interlinked with major challenges set down by biologists. Among the numerous examples of this "co-evolution" of sciences one can cite G.-L. Leclerc (1707-1788), known as *Comte de Buffon*, for his great work as both a naturalist and a mathematician. Recent technology breakthroughs have successively driven biology into the *genomic* and *post-genomic eras*. This quantum leap revealed the high complexity of biological organisms. Consequently the numerous and heterogeneous data produced every day require novel and efficient analysis methods for the biologists to investigate new hypotheses.

In 1984, Edwin Diday introduced the Pyramidal classification (Diday (1984)). It was one of the first methods that allowed determining and representing nested overlapping clusters. This approach became fully operational in 1990 with the publication of the complete ascending pyramidal classification algorithm (Bertrand (1990)).

The aim of this paper is to point out the potentiality of pyramids for the analysis of biological data. We present three applications dealing with genomic and transcriptomic data analysis. This examples illustrate that the inherent pyramid properties of overlapping and partial ordering can help with the interpretation of data.

This paper is organized as follows: first, two applications of pyramidal clustering are discussed on genomic data. One concerns genome scale sequence analysis, the other, the computation of multiple sequences alignments;

second, an application of pyramidal clustering is described with transcriptomic data obtained by DNA chips.

2 Genomic data

2.1 Genome scale sequence analysis

For several years, the success of numerous sequencing projects and their applications (*e.g.* transcriptom analysis) has led to the exponential increase of biological data. Thus, the availability of different genomes brought about the need for comparisons. For instance, by comparing the human genome with the genomes of different organisms, researchers can better grasp the structure and function of human genes and thereby develop new strategies in the battle against human diseases. In addition, comparative genomic (Konning et al. (1997), Park and Teichmann (1998)) provides a powerful new tool for the study of evolutionary changes among organisms, helping to identify genes that are conserved among species and genes giving each organism its own unique characteristics. In the context of comparative genomic, and among other methods, the pyramidal classification provided new interesting results (Codani et al. (1999), Aude et al. (1999)). More precisely, it allowed us to improve the representation and the analysis of the biological data. This point is fundamental: for example, it allowed to decipher the domain structure (functional subunit) of genes and to annotate genes (Louis et al. (2001)).

The following example deals with data from PHYTOPROT (Louis (2001)). This database is dedicated to the study of plants proteomes in order to elucidate functional relationships between genes of different species. All pairs of sequences have been compared and globally partitioned (Codani et al. (1999)); resulting clusters has been studied in details. Let study a family with the following proteins sequences: `APY_SOLTU` sequence of potato; `NTPA_PEA` sequence of garden pea; `O04519`, `O04520`, `O22204`, `O24091`, `O23505`, `O49676` sequences of *Arabidopsis thaliana*.

On figure 1.A, we have a dendogram obtained with the UPGMA clustering algorithm. We can observe two distinct clusters: the first one with the sequences `APY_SOLTU, NTPA_PEA, O49676`; the second one with all the others sequences. On figure 1.B, we have a pyramid computed on the same data. We rediscover both clusters, with an additional information. Indeed, the pyramid highlight the sequence `O49676` as a link between both clusters.

Then the domains decomposition of the sequences is computed using MKDOM (see figure 2). In the first cluster, the sequences `APY_SOLTU, NTPA_PEA` have all their domains in common and share one of them with sequence `O49676`. In addition `O49676` shares two domains with the sequences of the second cluster. Therefore, domains decomposition confirm that sequence `O49676` is a link between both clusters, as previously seen on the pyramid. The domain decomposition leads to the hypothesis that this sequence may be the

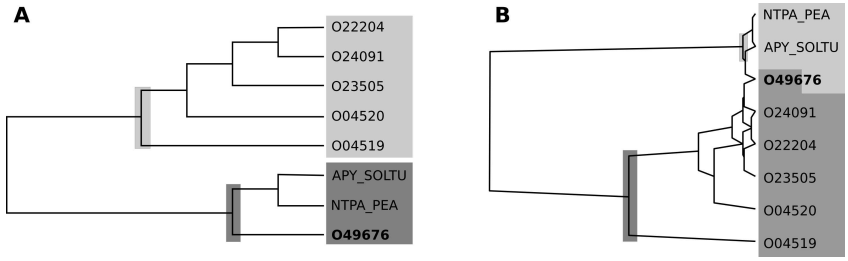


Fig. 1. A) The hierarchy obtained by the UPGMA method applied on a family of protein sequences from different plant organisms. One can unambiguously delineate two clusters, highlighted using grey boxes, from this hierarchy. **B)** The pyramidal representation obtained with the CAP algorithm on the same dataset. We observe two overlapping clusters, depicted by grey boxes. The intersection of both clusters is the sequence `O49676`. Thus, we can make the assumption that this sequence is the link between these two sets of sequences.

result of a gene fusion which is not detected by automatic syntactic annotation.

As a result, we can notice that the hierarchical representation is not able to determine links between two clusters. The pyramid with the properties of partial ordering and overlapping offers great interest for biological data. In this case, it permits to reconsider and correct the annotation of the sequence.

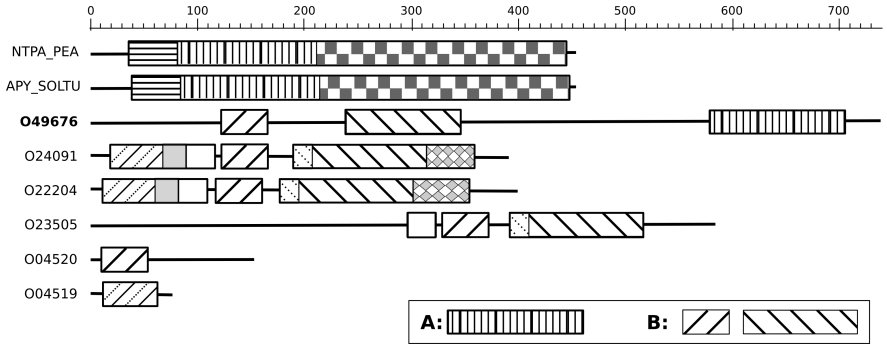


Fig. 2. This figure depicts the domains decomposition of protein sequences belonging to a PHYTOPROT family. They appear in the order given by the pyramid of figure 1.B. We observe that sequences of the first cluster (`APY_SOLTU, NTPA_PEA, O49676`) have the domain **A** in common, and the second cluster (all sequences from `O49676` to `O04519`) the domains **B**. The intersection of both clusters is the sequence `O49676`, which possesses both domains **A** and **B**. It demonstrates that this sequence links both sets of sequences. Moreover it is revealed by a clear visual diagram.

2.2 Multiple sequence alignments computations

Using a set of nucleotidic or peptidic sequences, one can try to identify conserved sequence regions among them. The classic method to discover such patterns is to compute a multiple sequence alignment (Feng and Doolittle (1987)). A multiple alignment arranges the sequences in a scheme where positions believed to be homologous are written in a common column. Like in a pairwise alignment, when a given sequence does not possess a nucleotide or amino acid in a particular position an insertion (denoted by a dash) is added. Multiple sequence alignment is certainly one of the most used method in bioinformatic, and researches in this area are still undergoing development (Batzoglou (2005)). In practice, it is a key step in various sequence analysis and covers a wide field of applications,including: sequence annotation (Bulyk (2003)); function and structure (secondary or tertiary) prediction (Jones (1999)); phylogenetic studies (Phillips et al. (2000)).

Among the numerous algorithms available to compute such alignments, a common strategy, called progressive, emerged from the vast majority of these methods. This strategy is made of three steps: (i) a similarity matrix is calculated using the scores of a pairwise alignment method applied on all possible pairs of sequences; (ii) this matrix is used to compute a hierarchical tree, usually named guiding tree; (iii) finally, the bottom-up exploration of this tree is used to select the pair of sequences (or a previously aligned subset of sequences) to align. All published progressive algorithm alter or refine one or more of these steps (Lee et al. (2002), Edgar (2004), Do et al. (2005), Katah et al. (2005)). Recently, Vescovo et al. (2005) has undertaken a study to estimate the impact of selecting other guiding structure, using alternative algorithms and parameters (*e.g.* neighbor-joining, hierarchical tree build using different aggregation criteria...), on the resulting alignment. Indeed, until now we have little knowledge about the effect of this tree on the final alignment.

Progressive alignments methods also differ in the way they compute each pairwise alignments within steps (i) and (iii). Some of them use global alignments (*e.g.* ClustalW, Thompson (1994)) in which sequences are aligned on their whole length. Others use local alignments (*e.g.* PIMA, Smith and Smith (1992)) in which only subsequences are optimally aligned. Recently a third way, usually called mixed, has been investigated that combined both global and local alignments (*e.g.* M-Align, Van Walle et al. (2004)). This new approach seems to achieve better results using standard benchmark databases (see Van Walle (2004)). Hereafter we will describe a new mixed progressive alignment algorithm that uses pyramidal clustering as a key component (Vescovo et al. (2004)).

This new method introduces some modifications in step (ii) and (iii) of the progressive strategy described above. In step (ii) the modification is straightforward. The guiding tree, usually computed using the neighbor-joining algorithm (Saitou and Nei (1987)) is replaced by a pyramid computed using

the CAP algorithm (Bertrand (1990)). The key idea is to use the overlapping properties of the pyramids to select the best alignment method (*i.e* global or local) in the step (iii). The principle of this algorithm is discussed with the example of sequences extracted from Thompson et al. (1994). The guiding structure is given in figure 3. Indeed, it makes sense to use local alignments when two set of sequences share a common pattern. This is precisely described by a cluster with a non empty intersection between its successors (*cf.* step 4 in figure 3). On the other hand, one can expect that successors with an empty intersection (*cf.* successors of step 5 in figure 3) don't reveal any shared pattern. In the latter case we would use global methods to align both sets of sequences. Moreover, the so-called local steps, such as step 4 in figure 3, require some adjustments in the definition of the two sets of sequences that are locally aligned. Basically, we have to deal with sequences that are present in each set, such as `HBB_HUMAN` in our example. We advocate to remove shared sequences from the largest set and to perform a local alignment. To preserve the key role of these shared sequences we also increase their weights, in the alignment procedure, to a significant extent.

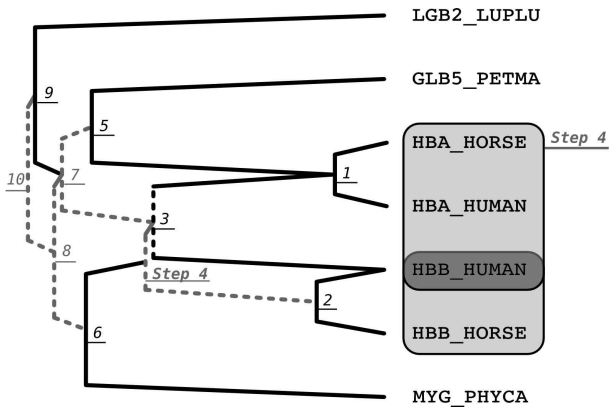


Fig. 3. This figure depicts a pyramid used as the guiding structure of a mixed progressive multiple sequence alignment algorithm. Labels on the right side of the figure are the swiss-prot accession numbers of the set of protein sequences to align. The pyramid is iteratively pruned by computing the consensus of the closest pair of sequences/consensus, according to the dissimilarity index (the steps numbers are indicated on each cluster). Solid lines indicate that the pair of sequences/consensus are aligned using a global method (steps 1, 2, 3, 5, 6, 9), whereas dashed lines indicate that sequences/consensus are aligned using a local algorithm (steps 4, 7, 8, 10).

We have successfully applied this new algorithm to the alignment of 11 homologous sequences from *Saccharomyces cerevisae* (*see* figure 4). All of them have been gathered in the same group using the genome scale analysis ap-

proach described in Codani et al. (1999) and previously explained. Querying the PFAM database (Bateman et al. (2004)), one can established that they have three domains in common: `Exo_endo_phos` (*exonuclease-endonuclease-phosphatase family*) depicted as black box on figure 4; `LRR` (*Leucine Rich Repeat*) as striped box; `PP2C` (*Protein Phosphatase 2C*) as grey box. But only `YAL021C` and `YJL005W` are composed of two distinct domains, respectively (`Exo_endo_phos`, `LRR`) and (`LRR`, `P2C`). Thus, a good multiple sequence alignment algorithm should not overlap these domains. To highlight the benefits of our strategy we have performed a comparison between three different softwares: ClustalW (Thompson et al. (1994)) the most used program to perform multiple sequence alignments that implements a global strategy; DiAlign (Morgenstern et al. (1996)) the standard local strategy method; PyrAlign (Vescovo et al. (2004)) the pyramid based mixed strategy described above.

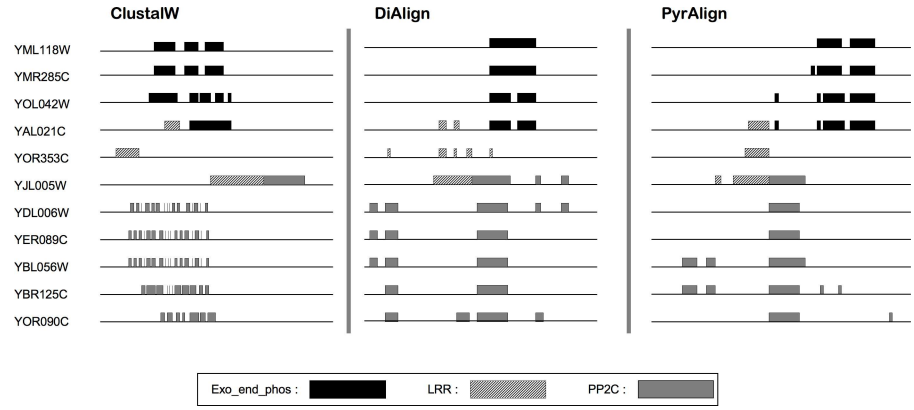


Fig. 4. This figure depicts the alignment of 11 sequences from *Saccharomyces cerevisae* computed by ClustalW, DiAlign and PyrAlign. The three domains, indicated within the box on the bottom of this figure, are used to benchmark these algorithms. ClustalW fails to correctly identify the domains: PP2C domain is split into many parts; LRR domain is not aligned across sequences; domains LRR and PP2C are overlapping. DiAlign also fails to build a correct alignment, domains are stacked and difficult to identify without supplementary knowledge. On the other hand, PyrAlign clearly delineates the domains and thus produces a better alignment.

On figure 4, one can easily notice that both ClustalW and DiAlign failed to split the three domains. Indeed, DiAlign stacks the `Exo_endo_phos` and `P2C` domains whereas ClustalW stacks all of them. PyrAlign is the only algorithm that clearly separate the domains, even if some of them are split in several parts such as `P2C`. Obviously all these programs fail to keep domains as continuous sequences of amino acids. For example ClustalW adds a lot of insertions within the `P2C` domain. DiAlign produces the same artifact when aligning the `LRR` domain. In one way, PyrAlign almost succeeds in keeping

domains as single blocks, but is less effective in delineating their borders (*e.g.* the left side of the `Exo_endo_phos` domain). We can also argue that PyrAlign splits several domains, but domain borders are fuzzy and heavily depend on the underlying algorithmic used to inferred them. For instance the `P2C` domain depicted on gene `YBR125C` is defined as a single block in the PFAM database and as three blocks in the Panther database (Paul et al. (2003)). This example demonstrates the efficiency of the PyrAlign algorithm and its pyramidal guiding structure. However, due to the highest number of clusters in pyramids, this method has to compute more alignments than the others. Consequently the complexity of this algorithm is higher than any other progressive method. This could be an issue if one wants to compute multiple alignments of large sets of sequences.

3 Transcriptomic data

In the mid 90's, the DNA chip technology (Schena et al.(1995)) made a breakthrough in analyzing gene expression on a genomic scale (*i.e.* the transcriptom). It allowed to quantify the activity of hundred of thousands of genes under various conditions of given cell extracts. Nowadays, after many improvements, DNA chips are daily used by biologists around the world. As a consequence, large amounts of data have been produced by this technology. For instance, the GEO database already collected millions of expression profiles for over 100 organisms, submitted by over 600 researchers (Barrett et al. (2005)). A drawback of this technology is that measures are usually very noisy. Therefore, exhibiting significant variations is a challenging task (see (Speed (2003)) for details). Once these genes detected, one usually search for delineating co-expressed sets of genes. In this context, numerous clustering methods have been used (Eisen et al. (1998)). In this section we will detail the advantages of using pyramids to analyze DNA chips.

Our motivations were to investigate the partial order, induced by the pyramidal clustering, to delineate co-expressed and co-localized genes in prokaryotes. Indeed, such a set of genes, called *operon*, is regulated by the same promoter and transcribed as single mRNA transcript. Because of their unique operon structure, prokaryotes offer an additional feature to decipher the global regulatory network under various conditions (*e.g.* oxidative stress, inactivation of transcription factors...). Unfortunately, automatic discovery of operons from the genome sequence is a difficult task and no universal method has emerged yet. Thus, new approaches forged on the integration of other useful informations, such as gene expression data, have been tried (Sabatti et al (2002)). In the latter, authors have used a Bayesian classification scheme to predict whether the genes are in an operon or not. Since genes in operons are transcribed at the same level, Carpentier et al. (2004) have used this property to benchmark several micro-array clustering methods on their capability to gather such genes. In this section we will show that pyramidal

classification is a very efficient method to discover sets of genes that are *potentially* transcribed as operon. Furthermore, pyramid graphs allow to easily identify co-expressed operon neighbors, providing a helpful tool to decipher regulatory mechanisms.

As part of the 2003-2006 French Nuclear Toxicology program (ToxNuc) we have been involved in the study of the effects of cadmium on several organisms. Cadmium and several cadmium-containing compounds are known carcinogens and can induce many types of cancer. This metal is used in many industrial processes such as metal plating and the production of nickel-cadmium batteries, pigments, plastics and other synthetics. Among the several organisms studied in this project, we focused our work on the cyanobacteria *Synechocystis*. *Synechocystis* is a unicellular non-nitrogen-fixing cyanobacterium and an inhabitant of fresh water. This organism has been one of the most popular organisms for genetic and physiological studies of photosynthesis. Our role in this project was to elucidate the molecular mechanisms involved in the cell response to cadmium toxicity. The transcriptom approach, using DNA chips, was used to characterize the kinetics of global changes in *Synechocystis* gene expression in response to continuous exposure to cadmium. Having processed all micro-arrays, we applied a linear model to exhibit significantly regulated genes. Then, we used a non-stringent p-value threshold ($p < 10^{-2}$), thus selecting ≈ 800 genes (*i.e.* the fourth of the entire genome). Finally a mixed hierarchical-pyramidal classification algorithm was designed to compare gene expression profiles based on their correlation. As a result we obtained a set of pyramids. The figure 5 is an excerpt of one of these pyramids that we will discuss hereafter.

Now we are able to check the ability of pyramidal clustering to efficiently report and predict operon genes. On figure 5 we have surrounded with grey box genes that are co-expressed and co-localized according to the pyramid. In addition we have checked that all genes of the same predicted operon are on the same DNA strain and oriented in the same direction. The first operon concerns genes involved in the motility of the cell. These proteins seem to be involved in bacteria fibrous proteins. These proteins are actually an operon (Yoshimura et al.(2002)) even if the gene `slr2018` is not annotated as a pilin-like protein. Furthermore these genes aren't neighbors within a hierarchy computed on the whole set of selected genes (data not shown). The second set of genes, predicted as an operon by the pyramid structure, reveal the efficiency of the method. On the figure we have manually annotated this operon as "hypothetical protein" because all corresponding genes have unknown functions according to Cyanobase (the cyanobacteria knowledge reference database). But mining other databases such as KEGG and the literature show that these genes are involved in the pilus assembly and required for mobility. Thus additionally to the fact that the method correctly predicts operon structures, it also gathers related operons. One more thing, the gene `sll1694` just below this operon is a known regulator of the pilus

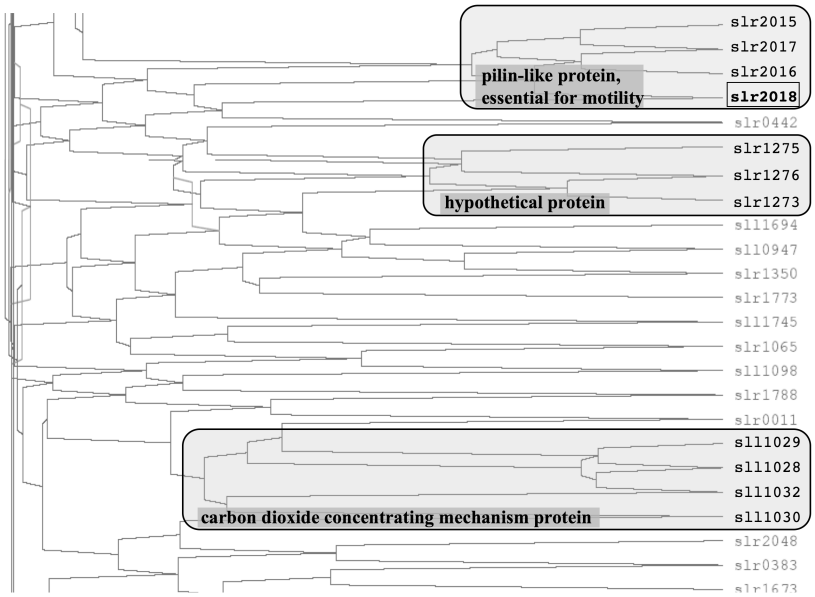


Fig. 5. This figure is an excerpt of the pyramidal classification of gene expression profiles from the cyanobacteria *Synechocystis*. Rounded boxes gather co-expressed and co-localized genes. These sets correspond to potential operons. Each of them is manually annotated using the Cyanobase database.

structure (Yoshihara et al.(2001)). Again this new element proved the accuracy of this approach. On the other hand `slr1274` has been missed, but one has to remember that these data are very noisy. The last predicted operon correctly gathered genes that are involved in the carbon dioxide concentrating mechanism. In this particular case our conclusions are motivated only by the similarity of genes annotations. Again, one gene `sll1031` is missing in this putative operon, for the same reasons as discussed previously.

In this section we have demonstrated the meaningful contribution of pyramidal clustering to the transcriptomic data analysis. This method should be considered with great interest for integrative approach of biological data analysis.

4 Discussion

In this article, we have shown the relevance of the pyramidal classification for biological data analysis. We illustrated this point through three different applications on *genomic* and *post-genomic* data. The first example discussed genome scale sequence analysis. It settled out the significance of clusters overlaps in deciphering links between families of proteins, thus improving sequences annotation. In the second example we used pyramids to specify a

new algorithm for computing multiple alignment of sequences. This method implements a mixed progressive approach that is very promising compared to standard algorithms. The last example is about transcriptomic data clustering using pyramidal classification. Here we demonstrate the potential of the partial order, induced by the pyramid, to identify *operons*.

Thus, perspectives of using pyramids for the analysis of biological data are very encouraging. Besides the examples given in this article, there are still many fields, in biology, to investigate using pyramids. But some issues, like the poor readability of pyramidal graphs, complicate its adoption by researchers. This may be solved by both improving the mathematical framework (Bertrand and Janowitz (2002)), and developping new suitable visualization systems. Finally, we will have to overcome minds for considering overlapping.

However, the *pyramid* concept is largely adopted by the biologists community. Indeed, MEDLINE, the life science bibliographic information repository, already indexes more than 500 articles with the word *pyramid* found in the title. Futhermore, it is interesting to notice that one of the main *systems biology* article is titled **Life's complexity pyramid** (Oltvai and Barabasi (2002)).

5 Acknowledgements

Authors thank Peggy Baudouin-Cornu at the Integrative Biology Laboratory (CEA, France) for critical reading.

References

AUDE, J.-C., DIAZ-LAZCOZ, Y., CODANI, J.-J. and RISLER, J.-L. (1999): Application of the pyramidal clustering method to biological objects. *Computer and Chemistry 23(3-4), 303-315.*

BARRETT, T., SUZEK, T.O., TROUP, D.B., WILHITE, S.E., NGAU, W.-C., LEDOUX, P., RUDNEV, D., LASH, A.E., FUJIBUCHI, W. and EDGAR R.(2005): NCBI GEO: mining millions of expression profiles – database and tools. *Nucleic Acids Research, Database issue 33, D562-D566.*

BATEMAN, A., COIN, L., DURBIN, R., FINN, R.D., HOLLICH, V., GRIFFTHS-JONES, S., KHANNA, A., MARSHALL, M., MOXON, S., SONNHAMMER, E.L.L., STUDHOLME, D.J., YEATS, C. and EDDY, S.R. (2004): The Pfam protein families database. *Nucleic Acids Research 32, 138-141.*

BATZOGLOU, S. (2005): The many faces of sequence alignment. *Briefings in Bioinformatics 6(1), 6-22.*

BERTRAND, P. and DIDAY, E. (1990): Une généralisation des arbres hiérarchiques : les représentations pyramidales. *Rev. Statistique Appliquée 38(3), 53-78.*

BERTRAND, P. and JANOWITZ, M.F. (2002): Pyramids and Weak Hierarchies in The Ordinal Model for Clustering. *Discrete Appl. Math., 122, 55-81.*

BULYK, M.L. (2003): Computational prediction of transcription-factor binding site locations. *Genome Biol., 5(1), 201.*

CARPENTIER, A.-S., RIVA, A., TISSEUR, P., DIDIER, G. and HENAUT A. (2004): The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA. *Comput Biol Chem. 28(1), 3-10.*

CODANI, J.-J., COMET, J.-P., AUDE, J.-C., GLEMET, E., WOZNIAK, A., RISLER, J.-L., HENAUT, A. and SLONIMSKI, P.P. (1999): Automatic analysis of large scale pairwise alignments of protein sequences. In: A.G. Craig and J.D. Hoheisel(Eds.): *Methods in Microbiology: Automation, Genomic and Functional Analysis.* Academic Press, (28) 229-244.

DIDAY, E. (1984): Une représentation visuelle des classes empiétantes : les pyramides. *INRIA, Rapport de Recherche No. 291.*

DO, C.B. and MAHABHASYAM, M.SP. and BRODNO, M. and BATZOGLOU, S. (2005): ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research 15, 330-340.*

EDGAR, R.C.(2004): MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32(5), 1792-1797.*

EISEN, M.B. , SPELLMAN, P.T., BROWN, P.O. and BOTSTEIN, D. (1998): Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A. 95(25), 14863-14868.*

FENG, D.F. and DOOLITTLE, R.F. (1987): Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution 25, 351-360.*

JONES, D.T.(1999): Protein Secondary Structure Prediction Based on position-specific Scoring Matrices. *J. Mol. Biol. 292, 195-202.*

KATOH, K., KUMA, K., TOH, H. and MIYATA, T. (2005): MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research 33(2), 511-518.*

KOONIN, E., MUSHEGIAN, A., GALPERIN M. and WALKER D. (1997): Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol Microbiol. 25, 619-637.*

LEE, C., GRASSO, C. and SHARLOW, M.F.(2002): Multiple sequence alignment using partial order graphs. *Bioinformatics 18(3), 452-464.*

LOUIS, A. (2001): La maitrise de l'information scientifique, clé de l'après séquencage *Thèse de l'Université Versailles Saint-Quentin.*

LOUIS, A., OLLIVIER, E., AUDE, J.-C. and RISLER, J.-L. (2001): Massive sequence comparisons as a help in annotating genomic sequences. *Genome Research 11, 1296-1303.*

MORGENSTERN, B., DRESS, A. and WERNER, T.(1996): DIALIGN: Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Nat. Acad. Sci. 32, 571-592.*

OLTVAI, Z.N. and BARABASI, A.L. (2002): Systems biology. Life's complexity pyramid. *Science 298(5594):763-4.*

PARK, J. and TEICHMANN, S. (1998): Divclus: an automatic method in the geanfammer package that finds homologous domains in single- and multi-domain proteins. *Bioinformatics 14, 144-150.*

PHILLIPS, A., JANIES, D. and WHEELER, W.(2000): Multiple sequence alignment in phylogenetic analysis. *Molecular Phylogenetics and Evolution 16(3), 317-330.*

SABATTI, C., ROHLIN, L., OH, M.K. and LIAO, J.C. (2002): Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res. 30(13), 2886-93.*

SAITOU, N. and NEI, M. (1987): The Neighbor-Joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution 4, 406-425.*

SCHENA, M., SHALON, D., DAVIS, R.W. and BROWN, P.O. (1995): Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science. 270(5235), 368-371.*

SMITH, R. F. and SMITH, T. F.(1992): Pattern-Induced Multi-sequence Alignment (PIMA) algorithm employing secondary structure-dependent gap-penalties for comparative protein modelling. *Protein Engineering 5, 35-41.*

SPEED, T. (2003): *Statistical Analysis of Gene Expression Microarray Data*. Chapman & Hall / CRC, Boca Raton FL.

THOMAS,P.D., CAMPBELL,M.J., KEJARIWAL, A., MI, H., KARLAK, B., DAVERMAN, R., DIEMER, K., MURUGANUJAN, A. and NARECHANIA, A. (2003): PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res. 13, 2129-2141 .Supplementary Materials.*

THOMPSON, J. D., HIGGINS, D.G. and GIBSON, T.J.(1994): Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research 22(22), 4673-4680.*

VAN MALLE, I., LASTERS, I. and WYNS, L.(2004): Align-m - a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics 20(9), 1428-1435.*

VESCOVO, L., AUDE, J–C., POLAILLON, G. and RislerJ–L.(2004): Progressive multiple alignment based on pyramidal classification and applied to multidomain proteins, *proceedings of the 12th International Conference on Intelligent Systems for Molecular Biology 2004, Glasgow, Scotland.*

VESCOVO, L., AUDE, J–C. and POLAILLON, G. (2005): Guide structure calculation: a critical step for the accuracy of progressive multiple sequence alignment algorithms. *proceedings of the 4th European Conference of Computational Biology 2005, Madrid, Espagne.*

YOSHIHARA, S., GENG, X., OKAMOTO, S., YURA, K., MURATA, T., GO, M., OHMORI, M. and IKEUCHI M. (2001): Mutational analysis of genes involved in pilus structure, motility and transformation competency in the unicellular motile cyanobacterium Synechocystis sp. PCC 6803. *Plant Cell Physiol. 42(1),63-73.*

YOSHIMURA, H., YANAGISAWA, S., KANEHISA, M. and OHMORI, M. (2002): Screening for the target gene of cyanobacterial cAMP receptor protein SYCRP1. *Molecular microbiology 43(4), 843-853.*

BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 14 2008, pages 1625–1631
doi:10.1093/bioinformatics/btn249

Systems biology

InteroPORC: automated inference of highly conserved protein interaction networks

Magali Michaut[1,2,*], Samuel Kerrien[2], Luisa Montecchi-Palazzi[2], Franck Chauvat[1], Corinne Cassier-Chauvat[1,3], Jean-Christophe Aude[1], Pierre Legrain[1] and Henning Hermjakob[2]

[1]CEA, IBITECS, Gif sur Yvette, F-91191, France, [2]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [3]CNRS, URA 2096, Gif sur Yvette, F-91191, France

Received on February 18, 2008; revised and accepted on May 26, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Protein–protein interaction networks provide insights into the relationships between the proteins of an organism thereby contributing to a better understanding of cellular processes. Nevertheless, large-scale interaction networks are available for only a few model organisms. Thus, interologs are useful for a systematic transfer of protein interaction networks between organisms. However, no standard tool is available so far for that purpose.

Results: In this study, we present an automated prediction tool developed for all sequenced genomes available in Integr8. We also have developed a second method to predict protein–protein interactions in the widely used cyanobacterium *Synechocystis*. Using these methods, we have constructed a new network of 8783 inferred interactions for *Synechocystis*.

Availability: InteroPORC is open-source, downloadable and usable through a web interface at http://biodev.extra.cea.fr/interoporc/

Contact: michaut.bioinfo@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Biological organisms live in complex interaction with their constantly fluctuating environment. Changes in regulatory networks have been observed in a number of organisms when they are under specific conditions or stressed by some environmental alterations, leading to modifications of their metabolism. The understanding of these phenomena depends not only on the knowledge of the numerous molecular effectors involved such as genes and proteins but also on the understanding of the functional relationships between them.

Experimental approaches used to decipher protein–protein interaction (PPI) networks are described in Shoemaker and Panchenko (2007a). To complement these experimental techniques, a number of computational methods have been developed to predict PPIs (Shoemaker and Panchenko, 2007b). Large-scale PPI networks are only available for a limited number of model organisms, thus systematic inference of PPIs has become a central task of functional genomics. Consequently, we have investigated network inference using the interolog concept originally introduced by Walhout *et al.* (2000) which combines known PPIs from one or more source species and orthology relationships between the source and target species to predict PPIs in the target species.

Since the original introduction of the interolog concept, such interactome transfers have been performed for different species and using different ortholog identification methods. Matthews *et al.* (2001) have transferred two large-scale, two-hybrid interaction maps of *Saccharomyces cerevisiae* onto *Caenorhabditis elegans*. PPI maps have been constructed for various organisms (Yu *et al.*, 2004) based on the *S.cerevisiae* interactome. Based on the InParanoid (Remm *et al.*, 2001) algorithm to identify orthologs, human networks have been inferred from several model organisms (Huang *et al.*, 2004, 2007; Lehner and Fraser, 2004; Persico *et al.*, 2005). Brown and Jurisica (2005) have developed the web-based database OPHID containing human PPIs using BLASTP and the reciprocal best hit approach (Jordan *et al.*, 2002). Maps have also been generated for *Plasmodium falciparum* (Wuchty and Ipsaro, 2007) or *Helicobacter pylori* (Wojcik *et al.*, 2002).

Such transfers have been done only for a limited number of species and no standard method or software seems to emerge. Each study was based on a combination of selected species and orthology computation methods. Yet a common tool usable for a large number of species would greatly facilitate comparative studies, leading to a better understanding of the extent of evolutionary conservation of PPI networks. Such a method would be of great help to decipher PPI networks in the wealth of organisms with a newly sequenced genome or still lacking identified PPIs. It usually takes several years to carry out genome-wide detection of PPIs. Consequently, we have developed an automated tool, interoPORC, to predict PPIs for all organisms present in the Integr8 database (Kersey *et al.*, 2005). This database systematically provides all deciphered genomes and their corresponding proteomes (655 organisms in release 75).

Through a multidisciplinary approach, we have investigated the biological responses to environmental stresses using the model cyanobacterium *Synechocystis* PCC6803. Cyanobacteria are the most abundant photosynthetic organisms on Earth and their living conditions are frequently challenged by changes in nutrient

*To whom correspondence should be addressed.

availability and exposure to pollutants. *Synechocystis* is a unicellular prokaryote with a small fully sequenced genome (3600 genes) (Kaneko *et al*., 1996) easily manipulable with replicating plasmids (Domain *et al*., 2004; Mazouni *et al*., 2004). It shares a wealth of homologous proteins with plants. Thus, lessons learned from stress responses in *Synechocystis* should greatly facilitate the understanding of how plants face environmental challenges. We used our interoPORC prediction tool based on orthologous protein clusters to predict PPIs for *Synechocystis*. In addition, we developed a second prediction method which was more flexible but required more computational resources. This method, called interoBH, was based on pairwise sequence comparisons. It was also applied to *Synechocystis*, starting with a limited set of source species. We selected several model organisms whose interactomes have already been investigated, namely *S.cerevisiae*, *Escherichia coli*, *Homo sapiens*, *Arabidopsis thaliana*, *C.elegans*, *Drosophila melanogaster* and *H.pylori*, which are representative of the overall biodiversity of living organisms. The use of both methods enabled us to construct a new network of 8783 PPIs for *Synechocystis*.

2 METHODS

2.1 Data sources

All selected genomes and proteomes were collected from Integr8 (Kersey *et al*., 2005). For *E.coli* and *H.pylori*, all the proteomes from the various sequenced strains were merged to generate a global proteome for each single species because PPIs are sometimes reported at the species rather than the strain taxonomic level. Furthermore, for proteins having multiple splice variants, we have only considered the longest product of the genes encoding them. The experimental PPI datasets from the three manually curated databases DIP (February 19, 2007) (Xenarios *et al*., 2002), IntAct (April 13, 2007) (Kerrien *et al*., 2007a) and MINT (April 5, 2007) (Chatr-Aryamontri *et al*., 2007), which provide tabular data files in the MITAB25 tabular format (Kerrien *et al*., 2007b), were downloaded. We merged all PPIs and removed duplications. Finally, we extracted physical interactions occurring in each of the seven species selected, without considering self-interactions. A total of 139 325 PPIs were included in our investigation (Table 1). We collected sequence similarities from the CluSTr database (Petryszak *et al*., 2005). The PORC orthology data were available from Integr8 (Kersey *et al*., 2005). The functional annotation of *Synechocystis* proteins is described in the GOA (Camon *et al*., 2004) file available from Integr8.

2.2 Prediction methods

For both prediction methods, we used the sequence similarity to identify putative orthologous proteins between species. Based on the interolog concept, we combined interaction datasets with orthology information to transfer PPIs from different species onto *Synechocystis*.

2.2.1 InteroBH We considered homologous proteins as putative orthologs between *Synechocystis* and each source organism. This method was called interoBH since it was based on a best hit approach. Homology predictions were derived from pairwise Smith–Waterman similarities with an *E*-value for each sequence comparison (Saebo *et al*., 2005). To select the best sequence homologies, sequence comparisons with an *E*-value less than 1E-10, a standard cutoff value, were considered (Martin *et al*., 2002; Yu *et al*., 2004). For each protein, we selected in each of the other organisms the best matching sequence as a homolog. In addition, if the former protein was the best matching sequence of another protein in the same species, we added the latter as another homolog. In this way, we modified the reciprocal best-hit approach in such a way that a given protein could have several homologs, considered as putative orthologs, in a single organism.

We then investigated each species separately. Let us take a transfer of *S.cerevisiae* interactome onto *Synechocystis* as an example. For each binary interaction occurring in yeast, we considered each interacting protein and looked for putative orthologs in *Synechocystis*. If both interacting proteins had a putative ortholog in *Synechocystis*, we transferred the PPI to these two putative orthologs. If a protein had several putative orthologs in *Synechocystis*, then we predicted all possible PPIs as putative PPIs. We used the joint *E*-value (Yu *et al*., 2004) to assess the quality of the predicted PPIs. The joint *E*-value was defined as the geometric mean of the individual *E*-values of both putative orthologs.

Table 1. Source interactions

Relevant organisms	*Proteins*	*Interactions*
S.cerevisiae	5780	54560
A.thaliana	758	1406
E.coli	3853	22023
H.sapiens	9234	26587
D.melanogaster	8636	27476
C.elegans	3275	5636
H.pylori	783	1637

For each organism, the number of interacting proteins and the number of PPIs collected in the databases IntAct, MINT and DIP are indicated.

2.2.2 InteroPORC InteroBH was generalized leading to a new method called interoPORC since it was based on the new PORC data (putative orthologous clusters) defined as orthologous families from Integr8. These clusters are of paramount interest since, unlike previously defined clusters (Koonin *et al*., 1998), they contain all sequenced organisms (556 bacteria, 59 eukaryota and 50 archaea in the release 75). Each entry in PORC represents a cluster of genes grouped by the similarity of their longest protein product. We used 215 733 clusters, containing 1 548 235 proteins. According to the PORC construction process, a cluster contains at most a single protein from a given species and a protein can be assigned only to a single cluster. In other words, it is impossible to find several proteins of the same species in a single cluster. When sequence comparisons were insufficient, the PORC algorithm did not attempt to resolve potential ambiguity using phylogenetic trees or network comparison (Bandyopadhyay *et al*., 2006). The clusters were split according to the weakest sequence similarity in order to respect the 'one gene per species per cluster' rule (Kersey *et al*., 2005).

The inference process was similar to that of interoBH, orthologous groups of proteins were used instead of binary orthology relationships between two species. The process was broken down into two steps. In the first step, called up-casting, we abstracted PPIs onto orthologous cluster links. For a given source PPI, if both proteins belonged to a cluster, we constructed a link between these two clusters. In the second step, called down-casting, we projected these cluster links onto target species to predict new PPIs. Practically, for a given link between two clusters, if both clusters contained a protein from the target species, we inferred a PPI between these proteins unless this PPI had been used as a source PPI to construct the cluster link.

2.3 Supporting evidence analysis

In order to support some of the predicted PPIs, we explored the following approaches: (i) PPI explanation on the basis of interacting domain annotation; (ii) sharing of functional annotations for both interacting proteins; (iii) prediction by several species; (iv) identification of source PPI by several experimental techniques; (v) comparison with experimentally identified interactions.

2.3.1 Domain–domain interaction score Interacting domain annotation was used to identify the predicted PPIs associated with domain pairs indicative of true interactions. We retrieved the Pfam domain composition of all proteins of *Synechocystis* from UniprotKB (Bairoch *et al.*, 2005). Then we collected a list of domain–domain interactions (DDI) derived from iPFAM structures (Finn *et al.*, 2005). We combined both types of information to compute a list of *Synechocystis* proteins that potentially interact together based on DDIs. Since we wanted to favor domain pairs that rarely occur in all protein pairs, we defined a score S_P, calculated using Equation (1). It is noteworthy that 78% of *Synechocystis* proteins were annotated with a single domain. Nevertheless, when a protein had several domains, we calculated the PPI score as the maximum score of all possible domain pairs.

$$S_p = \underset{d \in D}{\text{Max}} \left(\frac{1}{\text{count}(d)} \right) \quad (1)$$

D is the set of domain pairs constructed with the domain lists of both proteins of the PPI, p and count(d) is the number of occurrences of this domain pair d in all protein pairs of *Synechocystis*. We generated a set of 5000 random PPIs (Supplementary Material). Given that 95% of the scores were below 0.5, we considered all scores above this threshold as highly relevant.

2.3.2 Common Gene Ontology annotation Since interacting proteins either share similar functions or operate in the same biological process (Huang *et al.*, 2007), we considered both molecular function (MF) and biological process (BP) ontologies of the Gene Ontology (GO) (Ashburner *et al.*, 2000) and calculated semantic similarities between both proteins using the measure defined in Lubovac *et al.* (2006). We generated a set of 5000 random PPIs. Since 95% of random PPIs had a similarity below 0.23 for both MF and BP ontologies (Supplementary Material), we used this cutoff to identify PPIs with high semantic similarities.

2.3.3 Conserved interologs Some interactions were predicted several times from different source species using the interoBH method, which led us to think that they were meaningful (Lehner and Fraser, 2004). Similarly, different source interactions enabled us to construct a unique link between two clusters during the interoPORC process. The PPIs predicted from several source species were thus more likely to be valid. We therefore extracted the interactions predicted by several organisms.

2.3.4 Multiple experimental identification methods We examined the different kinds of experimental techniques which have been used to identify each source interaction. All experimental identification methods have different weaknesses and biases (Hakes *et al.*, 2008; von Mering *et al.*, 2002). Nevertheless, when an interaction has been detected by different methods, it is more likely to be genuine. In the MITAB25 format, a detection method is associated with each interaction using the PSI controlled vocabularies (Kerrien *et al.*, 2007b). We defined groups of methods as all children terms of the following controlled vocabulary terms: MI:0401 (biochemical), MI:0090 (Y2H), MI:0013 (biophysical), MI:0428 (imaging), MI:0254 (genetic), MI:0255 (transcription), MI:0063 (prediction), MI:0362 (inference), MI:0686 (unspecified). The list of all terms and their associated group is available in Supplementary Material file 1. Some of them do not appear in the source interaction data for the specific source (e.g. unspecified, transcription, genetic).

2.3.5 Comparison with experimental data On the one hand, we identified the predicted interactions that were present in the source dataset constructed from IntAct, MINT or DIP from low-throughput experiments. On the other hand, we analyzed all predicted interactions that overlap with the large-scale study recently published (Sato *et al.*, 2007).

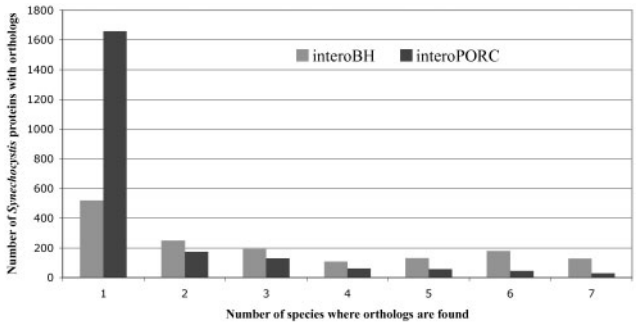


Fig. 1. Conservation of orthologous proteins in the seven selected species as predicted by each method.

3 RESULTS

To construct a PPI network for *Synechocystis*, we have developed two prediction methods, which combined known interactions from source species with putative orthology relationships. These two methods mainly differed in the way orthology relationships were constructed (see Methods), using either a pairwise approach based on best hits, interoBH or putative orthologous clusters, interoPORC. After showing some results about the orthology calculation, we will present the results of both prediction and interaction analysis methods, to end with the tool developed.

3.1 Orthology calculation

For each *Synechocystis* protein, putative orthologs were identified in the seven other species selected as reference organisms. We noted that the interoPORC method predicted more proteins with putative orthologs in only one species compared to the interoBH approach (Fig. 1). A smaller number of proteins were found to have putative orthologs across several species using interoPORC. This can be explained by the fact that the selected species are evolutionary distant not only from *Synechocystis* but also from each other. Thus only highly conserved proteins were found in a cluster with several of the selected species.

3.2 Interactions derived with InteroBH

We combined interaction dataset and orthology information to transfer interactions from seven source species onto *Synechocystis* separately (Table 2). Combining all results, we obtained a global set of 8586 interactions among 998 proteins (28% of the proteome of *Synechocystis*). This network was called interoBH_LOW. To assess the quality of the resulting interologs, we used the joint E-value defined in Yu *et al.* (2004). Since all sequence comparisons considered had an E-value less than the standard cutoff value of 1E-10, the joint E-value of each interolog was greater than 1E-10. Furthermore, it has been shown that a threshold of 1E-70 for the joint E-value enables a transfer of interactions with greater confidence (Yu *et al.*, 2004). Thus we considered all interologs with a joint E-value less than 1E-70 as a specific dataset called interoBH_HIGH. It should be noted that the orthology construction process used in the interoPORC method only considered sequence comparisons with a joint E-value less than 1E-40. Another dataset called interoBH_MEDIUM was considered for all interactions with a joint E-value less than this value.

Table 2. Number of PPIs predicted in *Synechocystis* by interoBH

	interoBH_HIGH		interoBH_MEDIUM		interoBH_LOW	
Source species	Inter	Prot	Inter	Prot	Inter	Prot
S.cerevisiae	955	299	1826	360	3558	438
A.thaliana	0	0	5	7	10	11
E.coli	1775	61	3183	744	4894	825
H.sapiens	26	26	69	74	194	150
D.melanogaster	14	16	30	37	97	95
C.elegans	1	2	3	6	21	35
H.pylori	199	75	164	117	251	160
Total	2870	1031	5280	1345	9025	1714
Non-redundant	2748	741	5070	884	8586	998

For each source species, the number of predicted interactions (Inter) and the number of proteins (Prot) involved in these predicted interactions are indicated. The Total line indicates the sum of all line values, whereas the Non-redundant line indicates the numbers of distinct interactions or proteins.

Not surprisingly, the number of predicted interactions was highly dependent on the number of available interactions in the source organism (Tables 1 and 2). It was also dependent on the evolutionary proximity to *Synechocystis*. Indeed, with almost the same number of source interactions, we transferred many more PPIs between the bacterium *E.coli* and the cyanobacterium *Synechocystis* than between *H.sapiens* and *Synechocystis*. It confirmed the recent result of (Brown and Jurisica, 2007) who showed that the number of interactions predicted by the interolog concept depends on the evolutionary distance between the organisms studied.

3.3 Interactions derived with InteroPORC

Using the interoPORC method, we predicted a dataset of 1446 interactions between 384 proteins in *Synechocystis*. In some cases, different source PPIs have been used to construct a single link between two clusters. In such a case, the predicted PPI was inferred from several source species.

3.4 Supporting evidence

In order to support some of the predicted interactions, we explored different methods based on interacting domain annotation, functional annotation, conservation across organisms, experimental techniques and experimentally identified interactions (see Methods).

3.4.1 Interacting domain annotation Within the union of interoPORC and interoBH_LOW, 177 interacting proteins shared a pair of domains from the set of known domain interactions. This set of PPIs included 39 associated with DDIs that had a highly relevant score (see Methods) increasing our confidence in these predicted interactions (Table 3). Itzhaki *et al.* (2006) have shown that DDIs frequently occur in protein complexes and are evolutionary conserved. Indeed, we observed some interconnected subgraphs representing complexes such as the RNA polymerase or the ATP synthase (Supplementary Fig. 1). Furthermore, Itzhaki *et al.* found that the number of PPIs explained by DDIs in the different PPI networks ranged from 6% to 20% only. Consequently, PPIs supported by DDIs are strengthened but PPIs not supported by DDIs are not necessarily weakened.

Table 3. Predicted PPIs associated with DDIs

Prediction sets	H	M	L	P
Total PPIs	2748	5070	8586	1446
PPIs with known domains	2689	4939	8197	1399
PPIs associated with DDI(s)	60	100	172	37
PPIs with highly relevant score	18	27	38	16

The DDI annotation is described for H: interoBH_HIGH, M: interoBH_MEDIUM, L: interoBH_LOW and P: interoPORC in terms of number of PPIs.

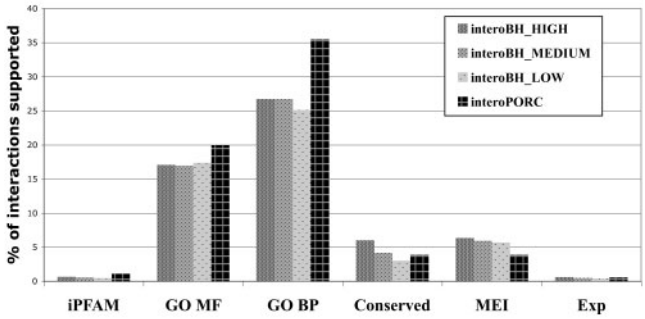


Fig. 2. Percentages of PPIs supported by different methods. iPFAM: PPIs explained by DDIs; GO MF (GO BP): interactions with similar terms in the MF (BP) ontology of GO; Conserved: interactions predicted by different source species; MEI: PPIs predicted from source PPIs identified by multiple experimental identification techniques; Exp: PPIs experimentally identified.

3.4.2 Functional annotation For the MF ontology, all networks derived with interoBH contained 17% of the interactions with similar annotation between their interacting proteins. The interoPORC network resulted in a slightly higher percentage (Fig. 2). For the BP ontology, interoBH networks resulted in about 26% of interactions with a similar annotation compared to 35% with the interoPORC network.

3.4.3 Conserved interologs The different datasets contained about 5% of conserved interologs (Fig. 2). We identified more conserved interologs in the interoBH_HIGH network. This was consistent with the fact that this network had been defined using a more stringent sequence comparison cutoff. There were seven highly conserved interactions transferred from three and four organisms. All 258 conserved interactions are represented in Supplementary Figure 2 between 167 proteins, including two highly connected chaperone proteins. This stressed the fact that interactions with a chaperone are detected with high-throughput identification techniques. This could be due to chaperone function that is used to bind to a number of proteins to assist their folding. However, non-specific binding cannot be ruled out. We noticed that 75% of all partners of the first chaperone groL1 (slr2076) do not share any GO term with it. These interactions were transferred from high-throughput detection methods such as two-hybrid or co-immunoprecipitation. We also examined existing data on the other highly connected chaperone protein dnaK2 (sll0170), and found that these predicted interactions were derived from interactions detected by different assays such as X-ray crystallography, molecular sieving, blue native

PAGE (polyacrylamide gel electrophoresis) or enzyme linked immunosorbent assay. This is consistent with the much higher rate of interacting partners sharing GO terms (40%).

3.4.4 Multiple experimental identification methods A total of 491 PPIs were transferred from source interactions identified by different experimental methods. The interoBH networks had 6% of interactions coming from several methods whereas the interoPORC network had 4% of such interactions (Fig. 2).

3.4.5 Comparison with experimental data Among all predicted interactions, 10 were among the 185 *Synechocystis* PPIs reported in the experimental datasets obtained from IntAct, MINT and DIP. To evaluate the significance of this overlap, we computed the probability of finding randomly an overlap greater than the one observed. We found according to a hypergeometric model that the probability was less than 1E-4 (Supplementary Material). Thus, the experimental results corroborated our predictions.

A further experimental study led to a new large-scale dataset of 3236 interactions between 1920 proteins (Sato *et al*., 2007). When we considered only the proteins included in this study, we had 3904 predicted PPIs instead of the 8783 PPIs predicted by interoPORC or interoBH_LOW. Among this predicted subset, Sato *et al.* identified 25 interactions, which was significant (P-value $<$1E-18). It is important to note that large-scale experimental datasets obtained with the same technique have an overlap smaller than 10% of the total number of interactions (Arifuzzaman *et al*., 2006), emphasizing the high false negative rate. We are currently investigating this comparison more in depth. Together, 35 predicted interactions have been experimentally identified (Supplementary Fig. 3).

Among the 8783 PPIs predicted by interoPORC or interoBH_LOW, we identified a core set of 3495 interactions supported either by interacting domain annotation, functional annotation, conservation across species, multiple experimental techniques or experimental identification (Supplementary Material file 2).

3.5 A tool of use for all sequenced genomes

Since the quality of the predictions depends on the quality of the source data, it was important to separate the prediction process and the source data used. We developed a stand-alone tool that can be applied to different source data, for example to high-quality PPIs and private datasets. In addition, interoPORC can be run on all platforms since it has been developed in Java (the source code is also available). Moreover, we have provided result files in standard formats (PSI25-XML and MITAB25) in order to interface easily with existing tools.

We also wanted to provide a tool that was fast and easy to use. Consequently, we have set up a web interface where predictions can be run just by giving a species identifier. We use source PPIs from IntAct, MINT and DIP as well as PORC data from Integr8. All source data are updated as soon as a new version is publicly available for any database.

As an illustration, we applied interoPORC to several representative organisms (Table 4). For example, we predicted 1678 interactions in the widely studied cyanobacterium *Anabaena* PCC7120, a model organism without any large-scale interaction dataset so far. All predicted PPIs are available on the web interface. First, we noted that we obtained 1% more PPIs for *Synechocystis*

Table 4. New PPIs for several organisms representative of the biodiversity predicted with interoPORC

Superkingdom	Species	*Taxid*	*Proteome*	*Curated*	*New*
Archaea	*P.kodakaraensis*	69014	2301	0	221
Archaea	*T.volcanium*	273116	1523	0	208
Eukaryota	*R.norvegicus*	10116	12028	2178	13469
Eukaryota	*A.fumigatus*	330879	9629	0	17225
Eukaryota	*P.falciparum*	36329	5283	2737	4026
Bacteria	*B.subtilis*	224308	4105	0	2160
Bacteria	*Synechocystis sp.*	1148	3506	185	1463
Bacteria	*Anabaena sp.*	103690	6070	1	1678

For each species, the superkingdom, the name (Species), the taxonomic identifier (taxid), the size of the proteome (Proteome), the number of PPIs in the source databases (Curated) and the number of new predicted PPIs (New) are indicated.

(1463 instead of 1446) than with the previous interoPORC prediction involving only the seven species with the largest PPI networks being involved (Table 3). Furthermore, the number of predicted PPIs was higher for eukaryota as compared to both archeae and bacteria. This can be due to the larger size of their proteome but also to the smaller evolutionary distance between source and target organisms since 83% of the source PPIs used occur in eukaryota (data not shown). This corroborates the results of Brown and Jurisica (2007) showing that the number of predicted PPIs decreases as the evolutionary distance increases.

Consequently, interoPORC is of great interest for every organism with a newly sequenced genome for which no large-scale interactome has been determined yet. It provides a raw picture of possible PPIs, which can be experimentally validated. For most species, a global PPI dataset is yet to be determined, emphasizing the value of this tool that quickly and easily gives new insights into PPI networks in a large number of organisms.

3.6 Discussion

In this study we have developed two new prediction methods, interoPORC and interoBH to infer PPIs. They are based on the interolog concept, combining source PPIs in several species with orthology relationships. The interoPORC method can be used to predict PPIs in the ever-increasing number of organisms with a newly sequenced genome where large-scale analyses remained to be carried out. In these organisms, it is now possible to quickly get a raw picture of possible interactions using the open-source automated tool interoPORC which can be run through a web interface or downloaded for stand-alone use. Moreover, with the increasing availability of PPI networks, recent studies have shown the benefit of PPI network comparison across evolution (Kalaev *et al*., 2008). However, large PPI networks are available for only a few model organisms so far. Therefore, interoPORC is of great interest for constructing new networks, leading to improved comparative studies.

The interoBH datasets tended to contain the interoPORC dataset when the cutoff on the joint E-value decreased. The overlaps with the interoPORC amounted to 580 (40%), 1069 (74%) and 1249 (86%) interactions for interoBH_HIGH, interoBH_MEDIUM and interoBH_LOW, respectively. We noted that the two methods differed in the way orthology was calculated. Several putative orthologs were detected with the interoBH approach while only one

protein was selected as a putative ortholog with the interoPORC method. To understand better the differences between the two approaches and assess to what extent the results were comparable, we investigated further the interoBH approach using the common reciprocal best-hit approach (Jordan *et al*., 2002), noted as interoRBH. Only one protein could be selected as a putative ortholog akin to the interoPORC approach. The proportion of interactions predicted only by the interoBH_MEDIUM method was highly reduced when considering interoRBH_MEDIUM, whereas the intersection between interoRBH_MEDIUM and interoPORC was only slightly reduced compared to interoBH_MEDIUM and interoPORC (data not shown). This confirmed that the additional interactions predicted by interoBH came from the choice to keep several putative orthologs. Moreover the interoPORC interactions have a lower joint *E*-value than the interoBH_LOW interactions (*P*-value <0.008, Supplementary Material). Consequently, the interoPORC method can be seen as a way to obtain a highly conserved interaction dataset.

It is worth noting that some interactions predicted by interoBH but not by interoRBH have been experimentally observed (Sato *et al*., 2007) and are thus relevant. Nevertheless, interoBH led to a higher number of predicted interactions than interoRBH. Thus we may expect a higher number of false positives as previously discussed in Yu *et al*. (2004). The interoPORC method proved to be a more stringent automated approach for all sequenced organisms. This raises the question of the tradeoff between the general and automatic nature versus the coverage and sensitivity of the different approaches. We propose here an automated tool of use for all species and we completed these stringent results with a more sensitive method for the particular species we were investigating.

The combined use of interoPORC and interoBH_LOW, enabled us to predict a global new network of 8783 PPIs for *Synechocystis* among which 3495 have been supported by different methods. Among these, 25 predicted PPIs have been identified in a new recently published large-scale dataset (Sato *et al*., 2007). Both experimental and computational approaches have weaknesses and miss lots of interactions. Thus it is highly interesting to have such computational methods at one's disposal to complete experimental datasets and identify interactions that may have escaped the experimental detection with high-throughput methods.

ACKNOWLEDGEMENTS

The authors thank Arnaud Martel for setting up the interoPORC web interface and Raphaël Guerois for insightful comments.

Funding: This work was funded by the European Commission (FELICS 021902 RII3) within the Research Infrastructure Action of the FP6 'Structuring the European Research Area' Programme (to S.K., L.M.-P., H.H.); French National Agency of Research (ANR Biosys06_134823 SULFIRHOM); European Commission, Marie Curie Fellowship (to M.M.); French Atomic Energy Commission grant (to M.M.).

Conflicts of interest: none declared.

REFERENCES

Arifuzzaman,M. *et al*. (2006) Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res.*, **16**, 686–691.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bairoch,A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**(Database issue), D154–D159.

Bandyopadhyay,S. *et al*. (2006) Systematic identification of functional orthologs based on protein network comparison. *Genome Res.*, **16**, 428–435.

Brown,K.R. and Jurisica,I. (2005) Online predicted human interaction database. *Bioinformatics*, **21**, 2076–2082.

Brown,K.R. and Jurisica,I. (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.*, **8**, R95.

Camon,E. *et al*. (2004) The Gene Ontology Annotation (GOA) Database–an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol.*, **4**, 5–6.

Chatr-Aryamontri,A. *et al*. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**(Database issue), D572–D574.

Domain,F. *et al*. (2004) Function and regulation of the cyanobacterial genes lexA, recA and ruvB: lexA is critical to the survival of cells facing inorganic carbon starvation. *Mol. Microbiol.*, **53**, 65–80.

Finn,R.D. *et al*. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

Hakes,L. *et al*. (2008) Protein-protein interaction networks and biology-what's the connection? *Nat. Biotechnol.*, **26**, 69–72.

Huang,T.W. *et al*. (2004) POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, **20**, 3273–3276.

Huang,T.W. *et al*. (2007) Reconstruction of human protein interolog network using evolutionary conserved network. *BMC Bioinformatics*, **8**, 152.

Itzhaki,Z. *et al*. (2006) Evolutionary conservation of domain-domain interactions. *Genome Biol.*, **7**, R125.

Jordan,I.K. *et al*. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.*, **12**, 962–968.

Kalaev,M. *et al*. (2008) NetworkBLAST: comparative analysis of protein networks. *Bioinformatics*, **24**, 594–596.

Kaneko,T. *et al.* (1996) Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.*, **3**, 109–136.

Kerrien,S. *et al.* (2007a) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**(Database issue), D561–D565.

Kerrien,S. *et al.* (2007b) Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, **5**, 44.

Kersey,P. *et al.* (2005) Integr8 and genome reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.*, **33**(Database issue), D297–D302.

Koonin,E.V. *et al*. (1998) Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.*, **8**, 355–363.

Lehner,B. and Fraser,A.G. (2004) A first-draft human protein-interaction map. *Genome Biol.*, **5**, R63.

Lubovac,Z. *et al*. (2006) Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins*, **64**, 948–959.

Martin,W. *et al*. (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl Acad. Sci. USA*, **99**, 12246–12251.

Matthews,L.R. *et al.* (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or 'interologs'. *Genome Res.*, **11**, 2120–2126.

Mazouni,K. *et al*. (2004) Molecular analysis of the key cytokinetic components of cyanobacteria: FtsZ, ZipN and MinCDE. *Mol. Microbiol.*, **52**, 1145–1158.

Persico,M. *et al*. (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, **6** (Suppl. 4), S21.

Petryszak,R. *et al*. (2005) The predictive power of the CluSTr database. *Bioinformatics*, **21**, 3604–3609.

Remm,M. *et al*. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.

Saebo,P.E. *et al*. (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res.*, **33**(Web Server issue), W535–W539.

Sato,S. *et al*. (2007) A large-scale protein protein interaction analysis in Synechocystis sp. PCC6803. *DNA Res.*, **14**, 207–216.

Shoemaker,B.A. and Panchenko,A.R. (2007a) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.*, **3**, e42.

Shoemaker,B.A. and Panchenko,A.R. (2007b) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, **3**, e43.

von Mering,C. *et al*. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

Walhout,A.J. *et al*. (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, **287**, 116–122.

Wojcik,J. *et al*. (2002) Prediction, assessment and validation of protein interaction maps in bacteria. *J. Mol. Biol.*, **323**, 763–770.

Wuchty,S. and Ipsaro,J.J. (2007) A draft of protein interactions in the malaria parasite P. falciparum. *J. Proteome Res.*, **6**, 1461–1470.

Xenarios,I. *et al*. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.

Yu,H. *et al*. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, **14**, 1107–1118.

Genome-wide location analysis reveals a role of TFIIS in RNA polymerase III transcription

Yad Ghavi-Helm, Magali Michaut, Joël Acker, Jean-Christophe Aude, Pierre Thuriaux, Michel Werner and Julie Soutourina

Genes & Dev. 2008 22: 1934-1947
Access the most recent version at doi:10.1101/gad.471908

Genome-wide location analysis reveals a role of TFIIS in RNA polymerase III transcription

Yad Ghavi-Helm, Magali Michaut, Joël Acker, Jean-Christophe Aude, Pierre Thuriaux, Michel Werner,[2] and Julie Soutourina[1]

CEA, iBiTec-S, Gif-sur-Yvette Cedex F-91191, France

TFIIS is a transcription elongation factor that stimulates transcript cleavage activity of arrested RNA polymerase II (Pol II). Recent studies revealed that TFIIS has also a role in Pol II transcription initiation. To improve our understanding of TFIIS function in vivo, we performed genome-wide location analysis of this factor. Under normal growth conditions, TFIIS was detected on Pol II-transcribed genes, and TFIIS occupancy was well correlated with that of Pol II, indicating that TFIIS recruitment is not restricted to NTP-depleted cells. Unexpectedly, TFIIS was also detected on almost all Pol III-transcribed genes. TFIIS and Pol III occupancies correlated well genome-wide on this novel class of targets. In vivo, some *dst1* mutants were partly defective in tRNA synthesis and showed a reduced Pol III occupancy at the restrictive temperature. In vitro transcription assays suggested that TFIIS may affect Pol III start site selection. These data provide strong in vivo and in vitro evidence in favor of a role of TFIIS as a general Pol III transcription factor.

[*Keywords*: TFIIS; transcription; RNA polymerase III; RNA polymerase II; ChIP–chip]

Supplemental material is available at http://www.genesdev.org.

Received January 21, 2008; revised version accepted May 23, 2008.

Gene transcription is a complex and highly regulated process that, in eukaryotes, is carried out by three specialized RNA polymerases (Pol I, II, and III), dedicated to the transcription of different sets of genes. Transcription starts by the assembly of large preinitiation complexes (PIC) comprising TBP (the TATA-binding protein common to all three transcription systems) and general transcription factors specific to the RNA polymerase considered. Other transcription factors facilitate RNA elongation by Pol II (Shilatifard et al. 2003). One of these elongation factors, TFIIS, was initially discovered in human (Natori et al. 1973) and yeast cells (Sawadogo et al. 1980b), and is highly conserved in the eukaryotic and archaeal kingdoms (Hausner et al. 2000; Fish and Kane 2002). Structurally unrelated but functionally equivalent factors (GreA and GreB) also operate in bacteria (Borukhov et al. 1993; Opalka et al. 2003). TFIIS forms a binary complex with Pol II and stimulates an intrinsic transcript cleavage activity of RNA polymerase allowing elongating enzymes to resume RNA synthesis after accidental transcription arrest (Fish and Kane 2002; Kettenberger et al. 2003).

TFIIS is organized in three domains as determined by limited proteolysis and nuclear magnetic resonance (NMR) (Awrey et al. 1997; Olmsted et al. 1998). The three-dimensional structure of the Pol II–TFIIS complex, without the TFIIS N-terminal domain, has been determined by crystallographic studies (Kettenberger et al. 2003). The N-terminal domain I of TFIIS protrudes from Pol II and is not required for elongation (Nakanishi et al. 1995; Awrey et al. 1998), but binds the yeast Mediator and SAGA coactivator (Wery et al. 2004). The TFIIS central domain II is inserted into the funnel-shaped pore of Pol II and brings the C-terminal conserved RSADE motif very close to the enzyme active site (Kettenberger et al. 2003). Remarkably, the C-terminal domain III of TFIIS shares the RSADE motif with the C-terminal parts of the Rpa12 (Pol I) and Rpc11 (Pol III) subunits, and is more distantly related to the corresponding Rpb9 subunit of Pol II, which has no RSADE motif. Rpc11 mediates the intrinsic transcription cleavage activity of Pol III (Chedin et al. 1998).

In the yeast *Saccharomyces cerevisiae*, TFIIS is encoded by *DST1*. The *dst1*-Δ mutant, or *dst1* mutants of the RSADE motif, grow like wild type, but are sensitive to the NTP-depleting effects of 6-azauracil (6AU) and mycophenolic acid (MPA), two drugs thought to compromise elongation efficiency (Exinger and Lacroute 1992; Ubukata et al. 2003). In the presence of 6-azauracil, deletion of *DST1* affects Pol II processivity without influencing its elongation rate (Mason and Struhl 2005). TFIIS was also

Corresponding authors
[1]E-MAIL julie.soutourina@cea.fr; FAX 33-1-69-08-47-12.
[2]E-MAIL michel.werner@cea.fr; FAX 33-1-69-08-47-12.
Article is online at http://www.genesdev.org/cgi/doi/10.1101/gad.471908.

found to stimulate transcription past an artificial arrest site in vivo, thus further supporting the notion that it overcomes elongation blocks (Kulish and Struhl 2001). Recent reports suggest that, in addition to its well-documented role as a Pol II elongation factor, TFIIS also contributes to transcription initiation (Malagon et al. 2004; Wery et al. 2004; Prather et al. 2005; Guglielmi et al. 2007; Kim et al. 2007). This initiation role does not depend on the C-terminal RSADE motif needed for transcript cleavage stimulatory activity (Guglielmi et al. 2007; Kim et al. 2007).

Although the role of TFIIS in stimulating Pol II transcript cleavage activity is well characterized in vitro, its in vivo function remains less obvious. In this study, we analyzed the genome-wide location of TFIIS using chromatin immunoprecipitation coupled with DNA microarray hybridization (ChIP–chip). TFIIS was detected on many Pol II-transcribed genes under normal growth conditions, and its occupancy closely correlated with that of Pol II. Unexpectedly, we identified a novel class of TFIIS targets corresponding to Pol III-transcribed genes. In vitro Pol III transcription assays suggested that TFIIS may contribute to correct start site selection. Taken together, our in vivo and in vitro data reveal a previously unsuspected role of TFIIS in Pol III transcription.

Results

TFIIS is located genome-wide on the Pol II- and Pol III-transcribed genes

Previous models proposed that TFIIS protein was recruited during transcription elongation in conditions where Pol II was stalled. This hypothesis was strengthened by ChIP assays, suggesting that TFIIS does not normally reside on DNA but is specifically recruited in the presence of NTP-depleting drugs that favor the transcriptional arrest of Pol II elongation complexes (Pokholok et al. 2002). However, recent ChIP analyses revealed the presence of TFIIS on promoters and coding regions of several selected genes in cells grown under normal conditions (Prather et al. 2005; Guglielmi et al. 2007). To better understand TFIIS function and resolve this discrepancy, we performed genome-wide location analysis of TFIIS under standard conditions, or in cells that have been NTP-depleted in the presence of MPA.

We examined the genome-wide distribution of TFIIS and Pol II by ChIP–chip experiments and compared their relative enrichments. The DNA arrays used contained >40,000 oligonucleotide probes covering 12 Mb of the yeast genome (see the Materials and Methods). This analysis was done using a strain (YGH2) carrying an active N-terminal HA-tagged version of TFIIS. We found 3652 oligonucleotides significantly bound (*P*-value <0.005) by TFIIS under standard conditions. Correlation between TFIIS and Pol II enrichment showed an unexpected pattern (Fig. 1A). The distribution was split into two highly different data sets. A first set of oligonucleotides, located within 1419 Pol II-transcribed genes, showed a good correlation between TFIIS and Pol II enrichment. As discussed below, a second group of oligonucleotides was enriched for TFIIS but not for Pol II, and corresponded to Pol III-transcribed genes.

The results obtained by ChIP–chip experiments were confirmed by conventional ChIP on a set of selected Pol II- and Pol III-transcribed genes (Fig. 1B). All selected genes displayed a significant enrichment compared to background level measured on the coding region of the nontranscribed *GAL1* gene. Similar results were obtained in ChIP experiments using chromatin from a strain expressing native TFIIS and anti-TFIIS antibodies for immunoprecipitation (data not shown). We wondered whether TFIIS could also be detected on Pol I-transcribed 35S rDNA and Pol III-transcribed 5S rDNA templates. Since the array used had a poor coverage of that region, we performed conventional ChIP experiments on the rDNA locus. A slight enrichment on the 35S rDNA and no significant enrichment on the 5S rDNA were seen compared to the intergenic NTS1 and NTS2 regions (Fig. 1C). Since the binding of TFIIS to transcribed sequences would also be observed if TFIIS was an RNA-binding protein, we tested if the enrichment of TFIIS was dependent on the presence of the transcribed RNA by digestion with RNase A/T1 in ChIP experiments. Digestion did not affect TFIIS enrichment (data not shown). We also examined TFIIS occupancy of a gene transcribed by a heterologous T7 RNA polymerase (Chen et al. 1987). We did not observe any significant TFIIS enrichment (Supplemental Fig. S1), indicating that TFIIS does not localize to transcribed sequences thanks to an RNA-binding property.

Genome-wide TFIIS occupancy on Pol II-transcribed genes correlates with Pol II occupancy

Analysis of Pol II genome-wide occupancy revealed 15,911 oligonucleotides significantly bound (*P*-value <0.005). These probes corresponded to 3819 Pol II-transcribed genes. After removing all the oligonucleotides corresponding to Pol III-transcribed genes from the analysis, there was a fairly linear relationship between TFIIS and Pol II enrichment with a correlation coefficient of 0.642 (Fig. 1A). The example of Pol II and TFIIS enrichment profiles on the *PYK1* gene (Fig. 1D) confirms that TFIIS is present on coding regions even under standard growth conditions (i.e., in the absence of NTP-depleting drugs), and that TFIIS and Pol II occupancies are extremely well correlated. Other examples of enrichment profiles on *ADH1* and *PGK1* genes are shown in Supplemental Figure S2A,B. The presence of TFIIS is not limited to transcribed protein-coding regions, since we detected TFIIS on intergenic regions transcribed in short unstable RNAs. The examples of the cryptic transcript *NEL025C*, located on chromosome V between *RMD6* and *DLD3* genes (Wyers et al. 2005), and of *SRG1*, a transcriptional repressor of *SER3* (Martens et al. 2004), are shown in Supplemental Figure S2C,D. Pol II- and TFIIS-enriched genes were classified according to the biological process categories of gene ontology (GO). A standard hypergeometric test was used to determine the overrepresented

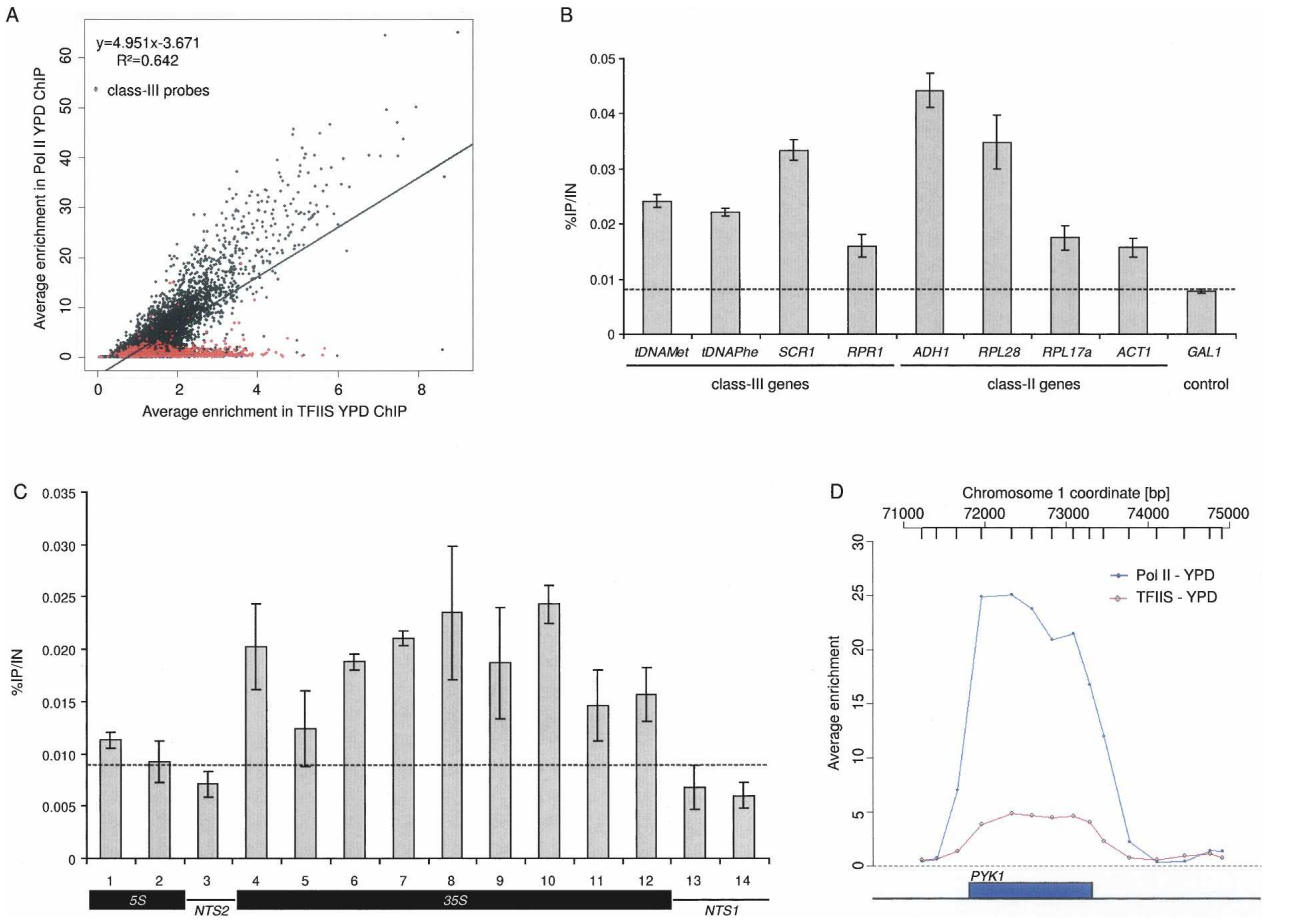


Figure 1. Genome-wide location analysis of TFIIS and Pol II. (*A*) Enrichment of Pol II versus enrichment of TFIIS. The genome-wide enrichment of Pol II and TFIIS in YPD medium was assessed from ChIP–chip experiments using YGH2 (3HA-TFIIS) strain. A linear regression for Pol II enrichment versus TFIIS enrichment on class II probes and its equation are indicated. Red dots correspond to class III probes. (*B*) Quantitative ChIP analysis of TFIIS enrichment on selected genes. Immunoprecipitated fragments from ChIP experiments were amplified with primers as indicated in the Supplemental Material. The *GAL1* ORF was used as a control. Error bars represent the standard deviation between at least three biological replicates. The background level was represented by a dotted line. (*C*) Quantitative ChIP analysis of TFIIS enrichment on rDNA. Immunoprecipitated fragments from ChIP experiments were amplified with primers as indicated in the Supplemental Material. The intergenic NTS1 and NTS2 regions were used as a control. Error bars represent the standard deviation between at least three biological replicates. The background level was represented by a dotted line. (*D*) Enrichment profile of Pol II and TFIIS on the *PYK1* gene. The enrichment of Pol II (blue) and TFIIS (red) on *PYK1* in YPD medium was assessed from ChIP–chip experiments using YGH2 (3HA-TFIIS) strain. The genomic positions of probe regions on chromosome 1 are indicated along the *X*-axis and represented by black points or circles. Watson strand-transcribed gene *PYK1* is colored in blue. The enrichment ratio is indicated along the *Y*-axis.

categories and associated *P*-values. Since the lists of Pol II and TFIIS overrepresented GO categories were largely overlapping (Supplemental Table S1), we concluded that TFIIS was not preferentially bound to a particular group of Pol II-transcribed genes. The smaller number of TFIIS-enriched genes compared to Pol II is best explained by the lower binding levels of TFIIS.

To explore the effect of NTP depletion on TFIIS genome-wide occupancy, yeast cells grown in minimal SD medium were exposed to 10 µg/mL MPA for 4 h (i.e., just when an effect on growth rate can be observed). We could detect a slight increase of TFIIS enrichment level compared to standard conditions (SD medium), but this effect was far less pronounced than in a previous study using a different and nonfunctional tagged version of TFIIS (Supplemental Fig. S3A; Pokholok et al. 2002). The enrichment profiles of TFIIS on *PYK1* and *ILV5* genes illustrate that in the presence of mycophenolate, occupancy levels of TFIIS were only slightly increased all along the genes (Supplemental Fig. S3B,C). Note that TFIIS enrichment is much higher in yeast cells growing rapidly in rich medium as compared with minimal medium.

Genome-wide TFIIS occupancy on Pol III-transcribed genes correlates with Pol III occupancy

We compared TFIIS and Pol III occupancies on Pol III-transcribed genes, since a global correlation would imply

that TFIIS might be a new Pol III general factor. A ChIP–chip experiment on Rpc160, the largest Pol III subunit, was performed. We found 964 enriched oligonucleotides located in or close to the class III genes that were described previously in genome-wide location analyses (Harismendy et al. 2003; Roberts et al. 2003; Moqtaderi and Struhl 2004). Four-hunderd-thirty-two of the 3652 oligonucleotides significantly bound (*P*-value <0.005) by TFIIS under standard conditions were within or close (up to 500 bp upstream or downstream) to class III genes. They were significantly enriched by Pol III, and represented all Pol III-transcribed genes for which probes were located on the arrays; i.e., *tDNAs*, *SCR1*, *RPR1*, *SNR52*, and *ZOD1*. *SNR6* showed a low level of TFIIS enrichment on the B box located ~100 bp downstream from transcribed region (Supplemental Fig. S4A). No TFIIS enrichment was detected on the *ETC* loci that are occupied by TFIIIC but not by TFIIIB and Pol III (Moqtaderi and Struhl 2004). As noted above, the 5S rDNA also showed no detectable enrichment for TFIIS.

The overall correlation between Pol III and TFIIS occupancies (Fig. 2A) confirmed our previous observations on the genome-wide TFIIS distribution. Considering only the oligonucleotides corresponding to Pol III-transcribed genes, the analysis showed a linear relationship with a correlation coefficient of 0.718 (Fig. 2A). The enrichment profile of TFIIS on *SCR1*, a selected class III gene, is presented in Figure 2B. Other examples of TFIIS enrichment profiles on *RPR1* and $tDNA^{Gly}$ are shown in Supplemental Figure S4B,C. In these examples, the Pol II-transcribed genes adjacent to class III genes were devoid of TFIIS, strongly suggesting that TFIIS could be recruited independently of Pol II transcription.

To establish that the presence of TFIIS on class III genes was specific to Pol III transcription and independent of Pol II transcription, we used specific Pol II and Pol III mutations and examined how they affected the distribution of TFIIS. The *rpb1-1* mutant of Pol II rapidly stops transcription after a shift to 37°C (Nonet et al. 1987). When cells were incubated at 37°C for 30 min, Pol II occupancy of all Pol II-transcribed genes tested (*ADH1*, *RPL28*, *RPL17a*, and *ACT1*) was greatly reduced in the *rpb1-1* strain compared to the wild-type strain (Fig. 3A, left panel). TFIIS occupancy strongly correlated with that of Pol II, and was also greatly reduced after incubation at 37°C. Conversely, TFIIS and Pol III occupancies of all Pol III-transcribed genes tested (*tDNAMet*, *tDNAPhe*, *SCR1*, and *RPR1*) remained largely unchanged under the same conditions (Fig. 3A, right panels). These results demonstrated that the presence of TFIIS on Pol III-transcribed genes is independent of Pol II transcription.

The Pol III-specific mutant *rpc25-S100P* was similarly used to impair class III gene transcription. For this purpose, a more prolonged shift to the restrictive temperature (10 h at 37°C) was needed (Zaros and Thuriaux 2005). Under this condition, Pol III and TFIIS occupancies were reduced on class III genes, even though the decrease of TFIIS occupancy was somewhat less pronounced (Fig. 3B, right panel). As expected, no significant effect was observed on Pol II and TFIIS association to the Pol II-transcribed genes (Fig. 3B, left panel). Thus, the presence of TFIIS on class II or III genes is largely dependent on transcription by Pol II or Pol III, respectively.

Most class III genes are too small to allow a spatial resolution of the Pol III machinery location by ChIP, but *SCR1*, the longest Pol III-transcribed gene (522 bp), provided this possibility. We examined the spatial distribution of TFIIS on this gene compared with that of the Pol III basal machinery composed, in addition to Pol III, of TFIIIB and TFIIIC complexes (Geiduschek and Kassavetis 2001). Seven real-time PCR amplicons spanning the upstream, transcribed, and downstream regions of

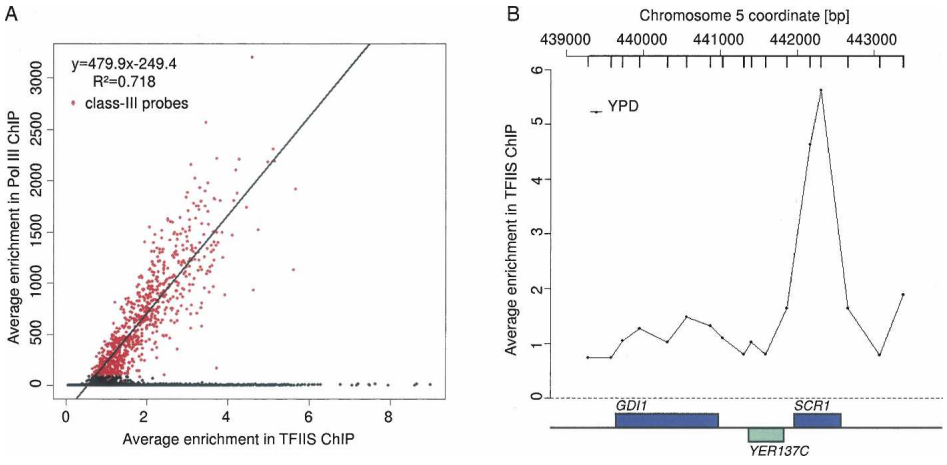


Figure 2. Genome-wide location analysis of TFIIS and Pol III. (*A*) Enrichment of Pol III versus enrichment of TFIIS. The genome-wide enrichment of Pol III and TFIIS in YPD medium was assessed from ChIP–chip experiments with MW671 (3HA-RPC160) and YGH2 (3HA-TFIIS) strains, respectively. A linear regression and its equation are indicated. Red dots correspond to class III probes. (*B*) Enrichment profile of TFIIS on the *SCR1* gene. The genome-wide enrichment of TFIIS in YPD medium was assessed from ChIP–chip experiments. The genomic positions of probe regions on chromosome 5 are reported along the *X*-axis and represented by black points. Watson strand-transcribed genes are colored in blue and Crick strand-transcribed genes are colored in green. The enrichment ratio is reported along the *Y*-axis.

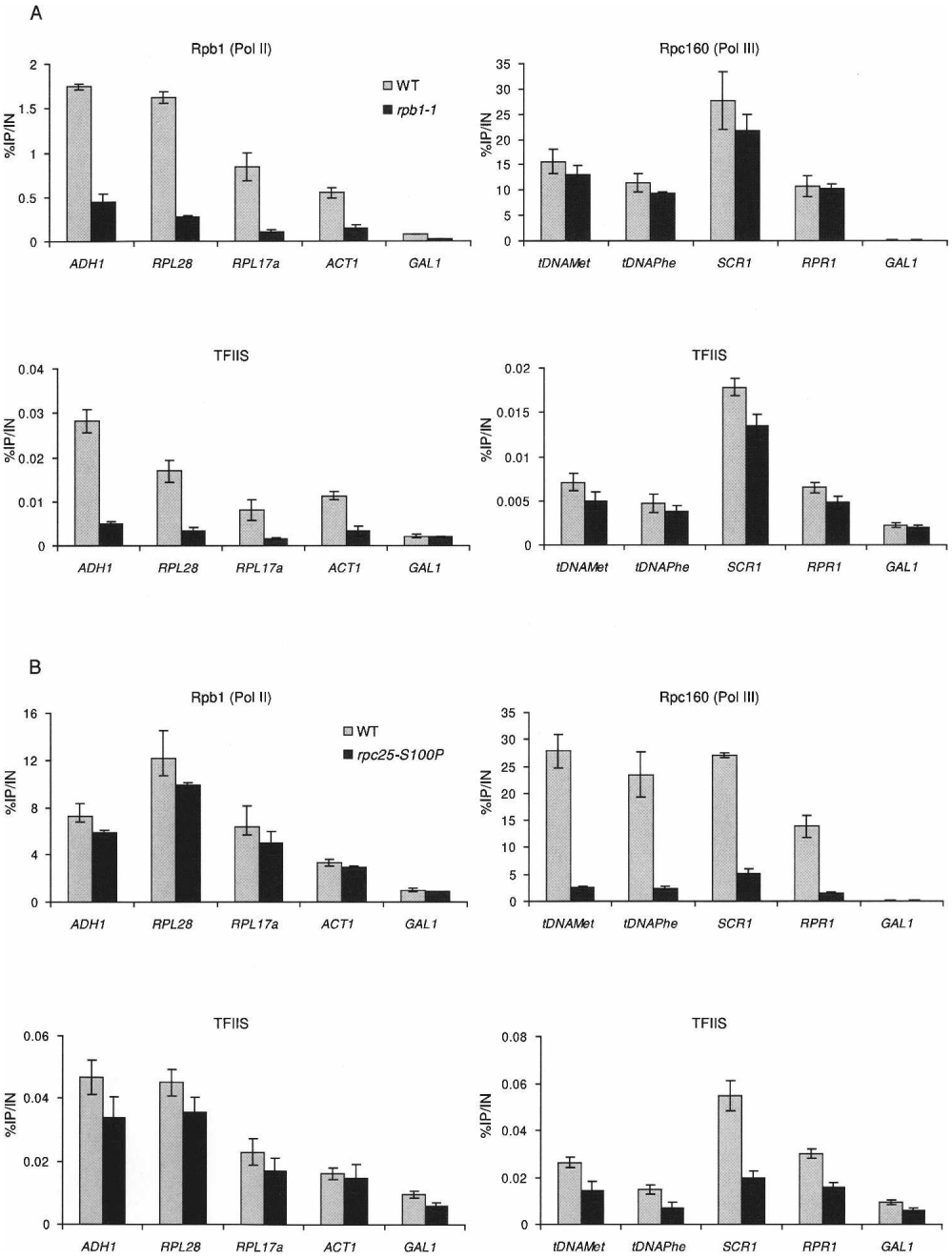


Figure 3. Effect of *rpb1-1* and *rpc25-S100P* mutations on TFIIS, Pol II, and Pol III occupancies on selected genes. Immunoprecipitations were performed using antibodies against 3HA (12CA5) for HA-TFIIS, 13Myc (9E10) for Rpc160-13Myc, and CTD (8WG16) for the Pol II Rpb1 subunit. Immunoprecipitated fragments from ChIP experiments were amplified with primers as indicated in the Supplemental Material. The *GAL1* ORF was used as a control. The values are the average of three independent experiments. Error bars indicate the standard deviation. (*A*) TFIIS, Pol II, and Pol III occupancies in the *rpb1-1* mutant. Standard ChIP assays were performed on chromatin prepared from the D788-4a strain transformed by *RPB1*-containing pYeB-B220 plasmid (wild type) or by empty vector Yep351 (*rpb1-1*). Cells were grown in selective SD medium complemented with amino acids at 30°C and then shifted for 30 min at 37°C. (*B*) TFIIS, Pol II, and Pol III occupancies in the *rpc25-S100P* mutant. Standard ChIP assays were performed on chromatin prepared from D792-3a strain transformed by *RPC25*-containing pRS315-RPC25 plasmid (wild type) or by empty vector pRS315 (*rpc25-S100P*). Cells were grown in selective SD medium complemented with amino acids at 30°C and then shifted for 10 h at 37°C.

SCR1 were designed (Fig. 4A). We analyzed the association profiles of the Rpc160 Pol III subunit, the Bdp1 TFIIIB subunit, the Tfc1 TFIIIC subunit, and TFIIS (Fig. 4B). The distribution of the Pol III transcription machinery on *SCR1* was as described previously (Roberts et al. 2006). TFIIIB binding was maximal on the TATA box,

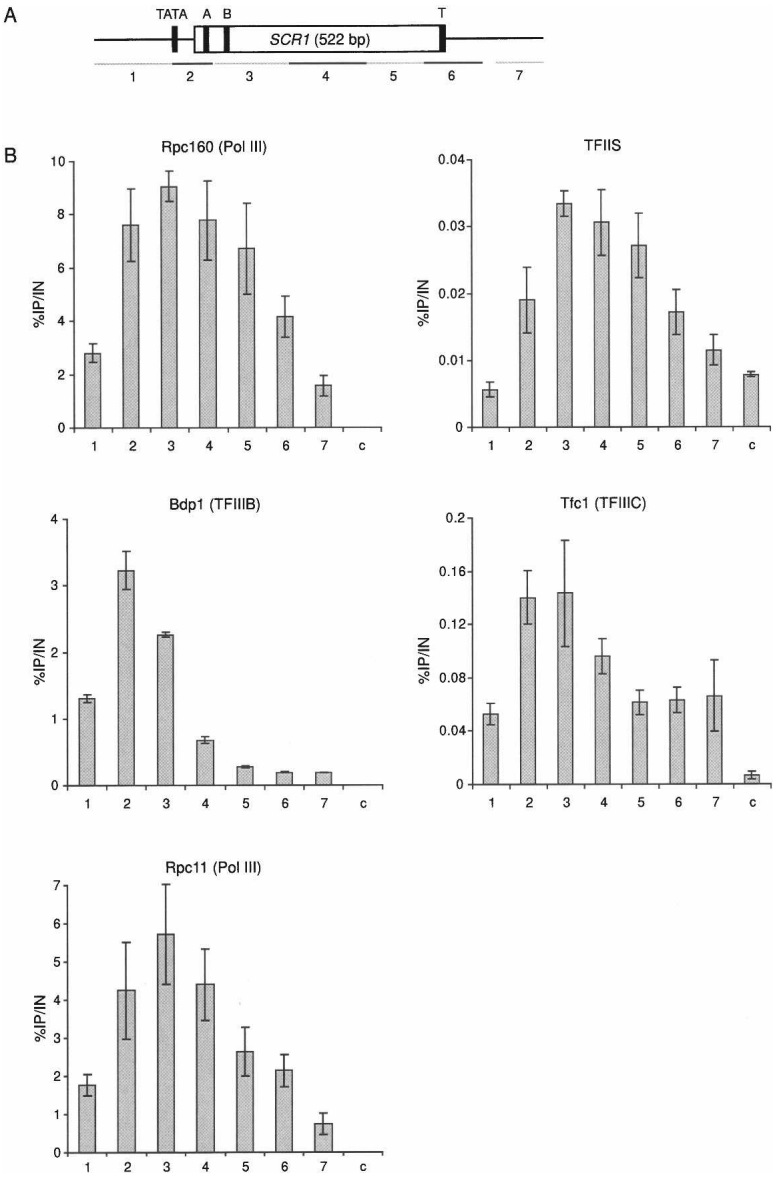


Figure 4. Pol III, TFIIS, TFIIIB, and TFIIIC location analysis on the *SCR1* gene. (*A*) Schematic organization of the *SCR1* gene. The location of the PCR fragments amplified in ChIP analyses are indicated by gray and black lines. Black boxes represent TATA, A and B boxes, and Terminator (T). (*B*) Occupancy profile of Pol III, TFIIS, TFIIIB, and TFIIIC on the *SCR1* gene. Standard ChIP assays were performed on chromatin prepared from YGH11 strain using antibodies against 3HA-TFIIS (12CA5) and Rpc160-13Myc (9E10), from MW4035 strain using antibodies against Bdp1-3HA (12CA5), from yOH1 strain using antibodies against Tfc1-13Myc (9E10), and from YGH15 strain using antibodies against Rpc11-3HA (12CA5). Cells were grown in YPD medium at 30°C. The *GAL1* ORF was used as a control (c).

while TFIIIC was cross-linked over the entire *SCR1* locus. TFIIS distribution on the *SCR1* gene resembled more closely that of Pol III with a maximum at the level of the A and B boxes at the beginning of the gene. However, the TFIIS occupancy profile differed from those of TFIIIC and TFIIIB. We also examined the association of Rpc11, a Pol III subunit homologous to TFIIS, and showed that the distribution of Rpc11 on the *SCR1* gene was similar to that of Rpc160 subunit (Fig. 4B). Thus, the TFIIS distribution on the *SCR1* gene follows closely that of Pol III, suggesting an active role of TFIIS in class III-gene transcription.

Since TFIIS interacts directly with Pol II, we tested the possible association of this factor with Pol III machinery by coimmunoprecipitation (co-IP) approach. No interaction of TFIIS with Pol III, TFIIIC, and TFIIIB was detected in crude extracts from noncross-linked cells (Supplemental Fig. S5A). However, when co-IP assays were performed with extracts from cross-linked cells (ChIP extracts), TFIIS coimmunoprecipitated with Pol III and TFIIIC, but not with TFIIIB (Supplemental Fig. S5B), in agreement with our ChIP data. We detected TFIIS when Rpc160-Myc or Rpc11-HA were used to immunoprecipitate Pol III. These co-IP results showed a concomitant presence of TFIIS with Pol III and TFIIIC on class III genes.

Effect of dst1 *mutations on tRNA synthesis and Pol III occupancy in vivo*

Since tRNA suppressor genes are transcribed by Pol III, we tested the influence of deleting *DST1* on suppression efficiency to analyze the role of TFIIS in Pol III transcription in vivo. We therefore compared the growth of a yeast strain containing the chromosomally integrated *SUP11* $tRNA^{Tyr}_{UAA}$ allele suppressing the *ade2-1* ochre mutation with that of the isogenic *dst1*-Δ strain. In the presence of adenine in the growth medium, the *SUP11*

dst1-Δ *ade2-1* strain grew as well as the *SUP11 DST1 ade2-1* strain (Fig. 5A, +Ade), but a significant growth difference was observed between the two strains in the medium lacking adenine (–Ade), indicating a decrease in the level of *ade2-1* mutation suppression. This observation suggests that TFIIS could influence Pol III transcription in vivo.

In the course of this study, we observed that *dst1*-Δ impaired yeast growth at 16°C (Fig. 5B), a phenotype that has not been described previously. This phenotype was not complemented by a point mutation in the RSADE motif of TFIIS domain III, responsible for the cleavage activity (*dst1-E291A*) (Fig. 5C). This mutant, the most sensitive to cold temperature, actually had a growth defect even at 30°C. In contrast, the *dst1-R200A* mutant in Pol II-binding domain II grew like a wild type. We concluded that the domain III of TFIIS is important for yeast growth at low temperature.

We analyzed the effect of *dst1* deletion on the levels of short-lived $tRNA_3^{Leu}$ precursor at low temperature. Wild-type strain and a *dst1*-Δ mutant were grown at 30°C or shifted for 8 h at 16°C. Total RNAs were extracted and used for Northern blot analysis (Fig. 5D). The amount of mature $tRNA_3^{Leu}$ was used as a loading control. The deletion of *dst1* significantly decreased the $tRNA_3^{Leu}$ precursor levels in yeast cells grown at 16°C (2.1-fold).

To investigate the effect of *dst1* mutations on class III gene expression under low-temperature conditions (16°C) in more details, we performed a genome-wide analysis of Pol III transcriptome (Ciesla et al. 2007). Log_2 ratios of class III gene transcription levels in the mutant versus

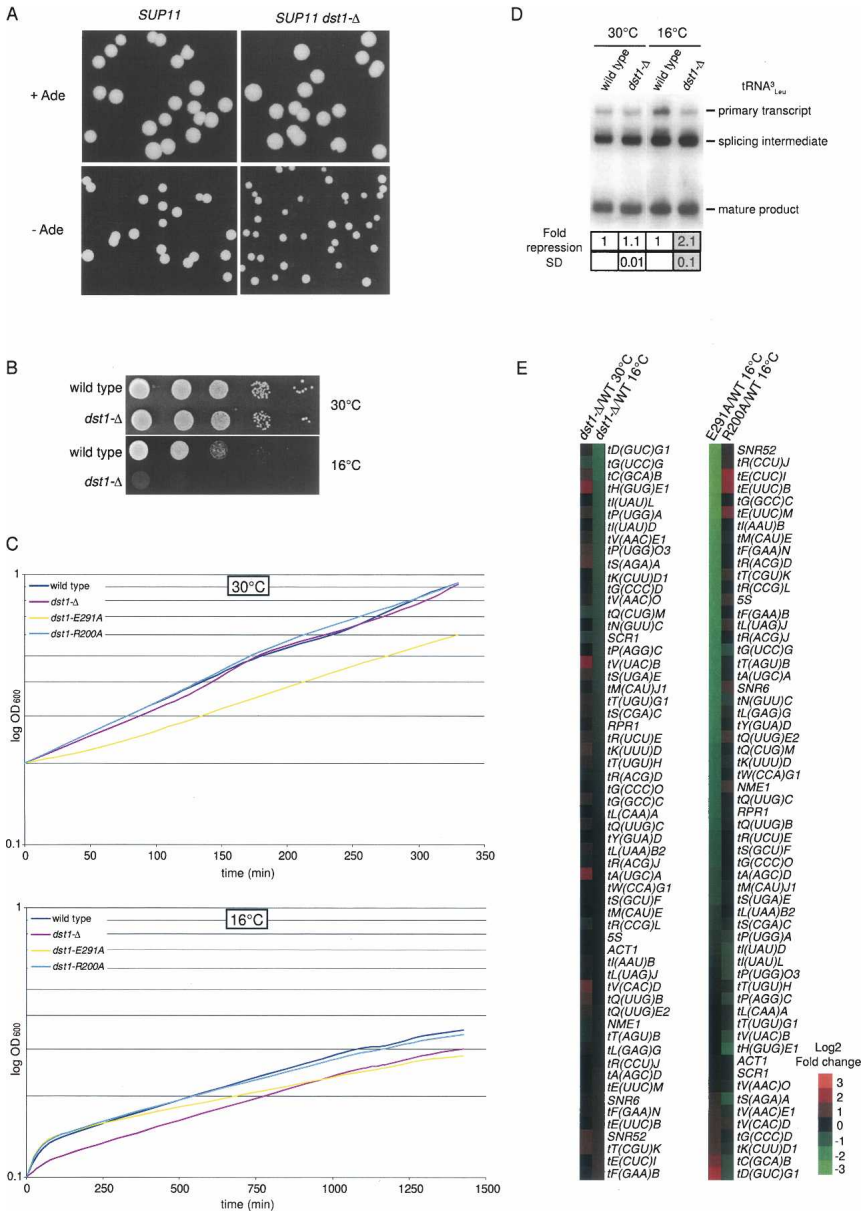


Figure 5. In vivo effect of TFIIS on Pol III transcription. (*A*) Growth of the YGH10 (*SUP11 dst1*-Δ *ade2-1 ochre*) strain relative to the MB159-2c (*SUP11 DST1 ade2-1 ochre*) strain. Cells were grown at 30°C for 3 d on minimal medium supplemented (+Ade) or not (–Ade) with adenine. (*B*) Growth of the *dst1*-null mutant at 30°C or 16°C. Suspensions of YPH500 (wild-type) or ESH1 (*dst1*-Δ) strains were serially diluted, spotted on YPD-rich medium, and grown for 3 d at 30°C or at for 5 d 16°C. (*C*) Growth of *dst1* mutants at 30°C or 16°C. *dst1*-Δ (YGH12) strain was transformed by pRS425-pPGK vector, pRS425-DST1, pRS425-dst1R200A, or pRS425-dst1E291A plasmids bearing different *DST1* alleles. Cells were grown in liquid SC-leucine medium at 30°C or shifted at time 0 at 16°C. (*D*) Northern blot analysis of $tRNA_3^{Leu}$ abundance in wild-type and *dst1*-Δ strains. RNAs were extracted from wild-type (ESH1/pRS425-DST1) and *dst1*-Δ (ESH1/pRS425-pPGK) cells grown in SC-leucine medium at 30°C or shifted at 16°C for 8 h. tRNA precursors abundance was quantified, and fold decrease of transcription in *dst1*-Δ strain was compared with that in wild type grown at the same temperature (arbitrary set to 1). Average values of three independent experiments and standard deviations (SD) are indicated. (*E*) Analysis of the Pol III transcriptome at 16°C. RNAs were extracted from the *dst1*-Δ (ESH1) strain transformed by pRS425-pPGK vector, pRS425-DST1, pRS425-dst1R200A, or pRS425-dst1E291A plasmids grown at 30°C or shifted for 8 h to 16°C, and analyzed by hybridization to Pol III-specific microarrays. Expression log_2 ratios of class III genes in the mutant strains compared with wild type are presented according to the red–green color scale.

wild-type strains were hierarchically clustered (Fig. 5E). At 30°C, the Pol III transcript profiles of the *dst1*-Δ and the wild-type strains were similar. In contrast, a significant decrease of several class III transcripts was observed in *dst1*-Δ mutant at 16°C. We found that transcription levels of class III genes were strongly diminished at 16°C in the *dst1-E291A* strain, whereas the *dst1-R200A* mutant had a Pol III transcript profile similar to that of the wild type. The *dst1-E291A* mutant had more pronounced effects on Pol III transcription than the *dst1*-Δ strain, consistent with its slower growth at all temperatures. Curiously, class III genes showing the most reduced transcript levels were different in *dst1*-Δ and *dst1-E291A* mutants.

To examine the effect of the *dst1*-Δ mutation on genome-wide Pol III occupancy at low temperature (16°C), wild-type and *dst1*-Δ strains containing a C-terminal 13Myc tag on the Rpc160 Pol III subunit were grown in YPD-rich medium at 30°C and then shifted to 16°C for 8 h. At low temperature, all class III genes were bound by Pol III. We compared Pol III genome-wide occupancy in a *dst1*-Δ and a wild-type strain grown at 16°C (Fig. 6A). A general reduction of Pol III occupancy in the *dst1*-Δ mutant was observed, suggesting that TFIIS could stabilize Pol III on class III genes. Regression analysis indicated a 1.5-fold decrease of Pol III binding in the mutant (the slope of the correlation line, shown in red, equal to 0.645) with a high correlation coefficient (R^2 = 0.965). The high correlation coefficient indicated that the association of Pol III to almost all class III genes was significantly reduced. A similar analysis was performed to examine the effect of *dst1* deletion on Pol II genome-wide occupancy at low temperature (Fig. 6B). The binding of Pol II was found to be reduced 1.6-fold in the mutant (the slope of the correlation line equal to 0.621) with a reduced correlation between wild-type and *dst1*-Δ strains (R^2 = 0.647). In the case of Pol II, the *dst1* deletion could have indirect effects on gene expression regulation that may explain the lower correlation coefficient between the wild-type and the mutant strains. Examples of enrichment profiles on specific genes are shown in Supplemental Figure S6. The overall reduction of Pol II and Pol III occupancies that was observed at low temperature when *dst1* was deleted suggested that TFIIS could stabilize Pol II and Pol III on their target genes.

We extended our genome-wide Pol III and Pol II location analysis to the *dst1-E291A* and *dst1-R200A* point mutants grown at 16°C. The binding of Pol II was found to be reduced 1.4-fold in the *dst1-E291A* mutant and 1.3-fold in the *dst1-R200A* mutant (the slope of the correlation lines equal to 0.723 and 0.788, respectively) (Fig. 6D,F). As expected, both *dst1* mutations affected Pol II occupancy. In contrast, the binding of Pol III was significantly diminished (3.1-fold) in the *dst1-E291A* mutant, but not at all in the *dst1-R200A* mutant (the slope of the correlation lines equal to 0.325 and 1.088, respectively) (Fig. 6C,E). Thus, the *dst1-E291A* mutant affected both Pol II and Pol III binding, whereas the *dst1-R200A* mutant had a small but significant effect only on Pol II association.

We further analyzed the occupancy of the Bdp1 TFIIIB subunit and the Tfc1 TFIIIC subunit in *dst1*-Δ and wild-type strains on several class III genes at 16°C (Fig. 6G,H). The association of TFIIIC was unchanged, but the binding of TFIIIB was reduced in the *dst1*-Δ mutant compared with the wild type.

TFIIS affects Pol III transcription in vitro

To examine the role of TFIIS in Pol III transcription in vitro, we expressed wild-type TFIIS and the TFIIS-E291A mutant form that is unable to stimulate Pol II elongation (Ubukata et al. 2003) as 6xHis fusion proteins in *Escherichia coli*, and purified the corresponding polypeptides to near homogeneity. We first checked that the wild-type TFIIS was active in stimulating nonspecific Pol II transcription on calf thymus DNA (Sawadogo et al. 1980b), while the TFIIS-E291A protein was not (data not shown). The effect of the wild-type and the E291A mutant TFIIS on Pol III transcription in vitro were examined in a Pol III transcription system reconstituted with all recombinant TFIIIC and TFIIIB and highly purified Pol III (Ducrot et al. 2006). Multiple-round transcription assays were performed with the *SUP4* tRNA gene as a template. As observed previously, the *SUP4* transcripts generated with TFIIIB recombinant components migrated as two or three diffuse bands on polyacrylamide gels (Fig. 7A, lane 1), which was not the case in the presence of the crude fraction B″ (Andrau and Werner 2001). This fraction contains additional factors, like Nhp6 (Braglia et al. 2007), that can restore transcriptional initiation specificity (Kassavetis and Steiner 2006). Western blotting analysis also revealed the presence of TFIIS in the B″ fraction (data not shown). Remarkably, adding the purified wild-type TFIIS to the reconstituted transcription system resulted in the formation of the correct length transcript (Fig. 7A, lanes 2–6), in contrast to the purified TFIIS-E291A mutant protein at the same concentration (Fig. 7A, lanes 7–11). The same results were obtained after mixing wild-type and mutant TFIIS preparations, indicating that the mutant TFIIS preparation did not contain interfering components (Fig. 7A, lane 12). Quantification of the transcription signals revealed that the total amount of transcripts did not significantly change in the presence of wild-type or E291A mutant TFIIS. High concentrations of TFIIS (40 ng/μL) started to inhibit the transcription (Fig. 7A, lanes 6–11). In the subsequent assay, we used 5 ng/μL TFIIS.

To demonstrate that TFIIS could contribute to correct start site selection, RNAs generated in a standard multiple-round Pol III transcription assay were purified and analyzed by primer extension. Figure 7B showed that, in the absence of TFIIS, transcription initiation occurred at base pairs +1, +4, and +8 relative to the start site used in vivo. The addition of B″ fraction or of wild-type TFIIS restored a correct start site selection, while the TFIIS-E291A mutant had no effect. Similar results were obtained on transcripts produced in a single-round assay (Supplemental Fig. S7A). The transcription start sites were analyzed in vivo on *SUP4* gene in a *nhp6a*-Δ *nhp6b*-Δ con-

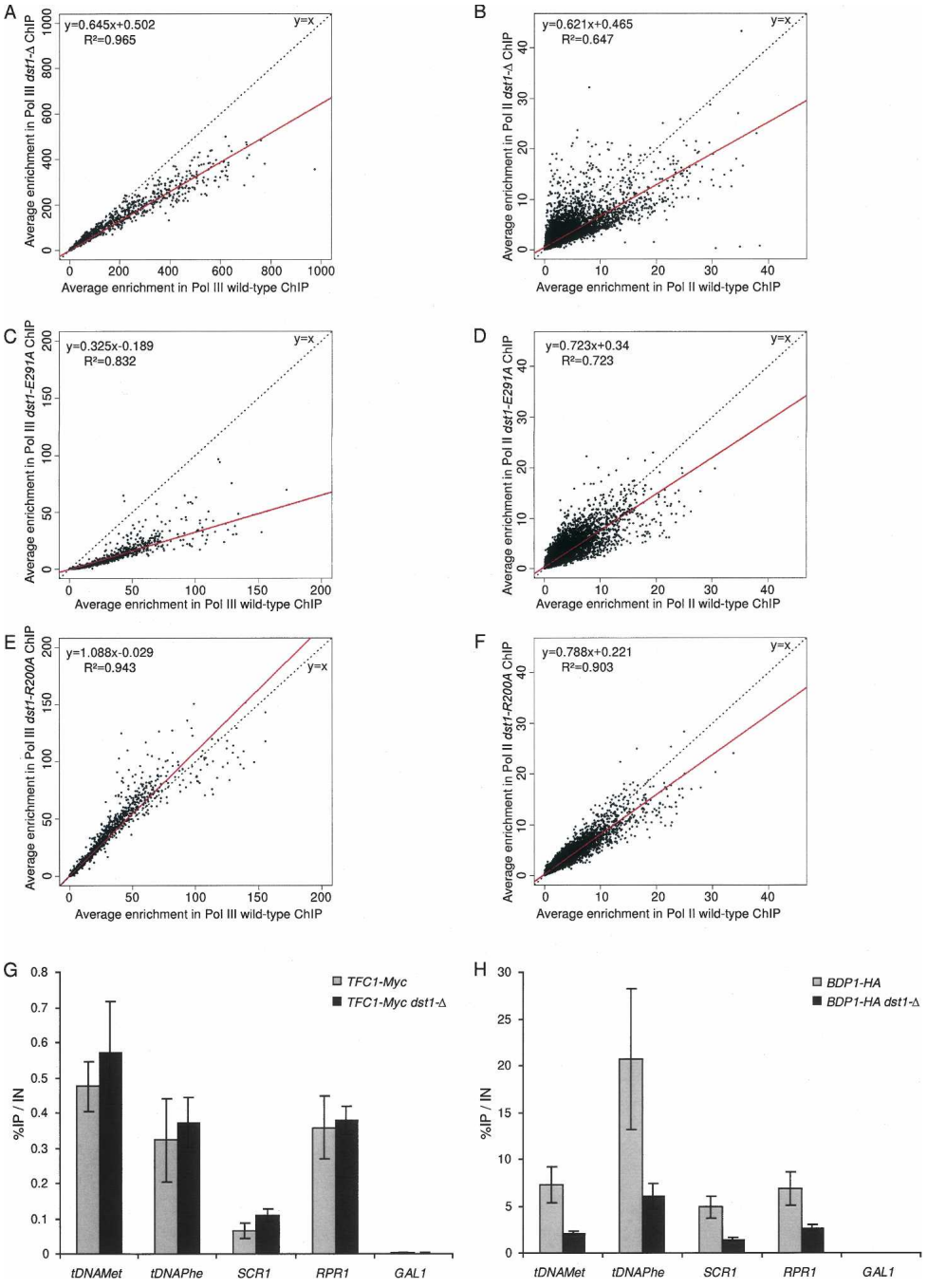


Figure 6. Pol II and Pol III genome-wide enrichment in TFIIS mutants at 16°C. The genome-wide enrichment of Pol III (*A*,*C*,*E*) or Pol II (*B*,*D*,*F*) in YPD medium (*A*,*B*) or SC-leucine medium (*C*–*F*) at 16°C was assessed from ChIP–chip experiments. A linear regression (red line) and its equation are indicated. The dotted line corresponds to $y = x$. Enrichment of Pol III or Pol II in *dst1*-Δ (YGH12) was compared with the wild-type strain (YGH11) (*A*,*B*). Enrichment of Pol III or Pol II in *dst1-E291A* (YGH12/pRS425-dst1E291A) (*C*,*D*) and *dst1-R200A* (YGH12/pRS425-dst1R200A) (*E*,*F*) strains was compared with the wild-type strain (YGH12/pRS425-DST1). (*G*,*H*) Quantitative ChIP analysis of TFIIIC and TFIIIB enrichment on selected class III genes. *dst1*-Δ and wild-type strains were grown in YPD medium at 30°C and shifted for 8 h to 16°C. Immunoprecipitations were performed using antibodies against 13Myc (9E10) for Tfc1-13Myc and 3HA (12CA5) for Bdp1-3HA. Immunoprecipitated fragments from ChIP experiments were amplified with primers as indicated in the Supplemental Material. The *GAL1* ORF was used as a control. Error bars represent the standard deviation between at least three biological replicates.

text and in the presence or absence of *DST1* gene. The *nhp6* mutant background was used because these proteins were previously implicated in start site selection and could mask the effect of *dst1*-Δ. We could not identify any effect of *dst1* deletion on start site selection in this background (Supplemental Fig. S7B).

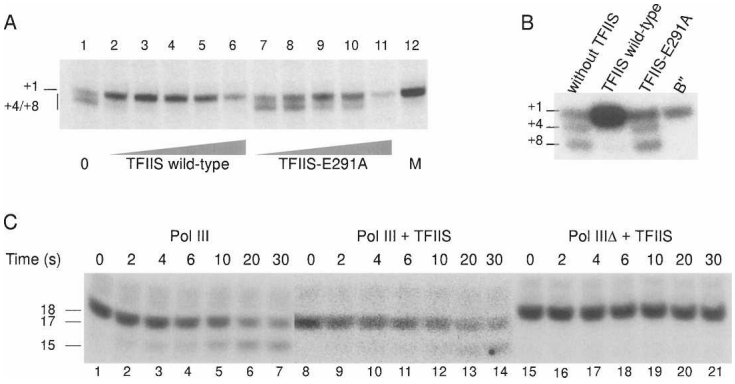


Figure 7. TFIIS stimulates faithful Pol III transcription in vitro. (*A*) Effect of TFIIS on Pol III transcription. Standard in vitro Pol III transcription on the *SUP4* template has been performed as described in the Materials and Methods, in the absence (lane *1*) or presence of increasing quantities (0.1, 0.2, 0.4, 0.8, and 1.6 µg) of wild-type TFIIS (lanes *2-6*) or mutant TFIIS–E291A (lanes *7–11*) or with a mix of wild-type and E291A protein (M, lane *12*). Transcription start sites are indicated. (*B*) Primer extension analysis of start sites used by Pol III on the *SUP4* gene in vitro. Transcription reactions were performed in the absence of TFIIS, in the presence of wild-type TFIIS, mutant TFIIS–E291A, or with a B″ fraction. Transcription and primer extensions were performed as described in Materials and Methods using a probe hybridizing within the *SUP4* intron. Positions of the major transcription start site are indicated. (*C*) Time-course analysis of RNA cleavage by Pol III. Pol III (lanes *1–14*) or Pol IIIΔ (lanes *15–21*) (Chedin et al. 1998) ternary complexes formed in the presence of 3XTPs were isolated on Sepharose CL-2B as described in the Supplemental Material and then incubated for 10 min with a 50 M excess of purified TFIIS (lanes *8–21*) or without TFIIS (lanes *1–7*). Ternary complexes were then incubated at 16°C in transcription buffer containing 5 mM $MgCl_2$ in the absence of nucleotides for various periods of time. The transcript sizes are indicated. Transcription with Pol IIIΔ results in the formation of an 18-mer RNA instead of a 17-mer for the wild-type enzyme (Chedin et al. 1998).

To examine whether TFIIS could affect the elongation step of Pol III transcription, the elongation kinetics of Pol III were analyzed on a *SUP4* template that can produce a stalled ternary complex after the synthesis of a 17-mer transcript in the absence of GTP. The ternary complex was allowed to resume transcription by adding the four NTPs. No significant changes in elongation kinetics were observed in the presence or absence of purified wild-type TFIIS (Supplemental Fig. S8). TFIIS being a cleavage stimulatory factor in Pol II transcription, we wanted to know whether TFIIS could influence the intrinsic cleavage activity of Pol III. Labeled ternary complexes halted by omission of GTP in the transcription reaction were purified and incubated for various periods of time in the presence of $MgCl_2$ to activate the Pol III cleavage activity (Chedin et al. 1998). TFIIS did not stimulate the cleavage activity of wild-type Pol III (Fig. 7C, lanes 1–14, look at the disappearance rate of the 17-mer transcript). Pol IIIΔ, a RNA polymerase mutant that lacks the Rpc11, Rpc37, and Rpc53 subunits, is not competent for RNA cleavage activity (Chedin et al. 1998; Landrieux et al. 2006). Pol IIIΔ was purified from *rpc37HA-Ct* mutant as described previously (Landrieux et al. 2006). We observed that the addition of TFIIS did not restore an efficient cleavage by Pol IIIΔ (Fig. 7C, lanes 15–21). Taken at face value, in vitro transcription assays suggest a role for TFIIS in start site selection during Pol III transcription.

Discussion

In this study, we performed a genome-wide location analysis of the TFIIS transcription factor. TFIIS was detected across the whole genome of exponentially growing cells, indicating that the binding of this transcription factor to chromatin is not restricted to NTP-depleted cells. TFIIS and Pol II genome-wide occupancies correlated very well, suggesting that TFIIS is not recruited only when Pol II is stalled. A second and more surprising outcome of our study was that TFIIS could be detected on almost all Pol III-transcribed genes. This result raised the intriguing possibility that TFIIS might operate as a general Pol III-associated factor. We provided substantial in vivo and in vitro data demonstrating that TFIIS is important for Pol III transcription.

Concerning the TFIIS function in Pol II transcription, we observed that TFIIS was associated with a large number of Pol II-transcribed genes in cells growing exponentially in rich medium, in line with previous ChIP analyses of TFIIS on a few class II genes (Prather et al. 2005; Guglielmi et al. 2007). The distribution of TFIIS over distinct Pol II-transcribed regions, including intergenic regions transcribed in short unstable RNAs (Martens et al. 2004; Wyers et al. 2005; Davis and Ares 2006), precisely correlated with Pol II itself, which is consistent with the role of TFIIS as an elongation factor but is not contradictory with an additional role during initiation (Prather et al. 2005; Guglielmi et al. 2007; Kim et al. 2007). In the presence of MPA, the enrichment level of TFIIS genome-wide was only slightly increased compared with that in normal growth condition. Essentially, two models for TFIIS recruitment to Pol II-transcribed genes may be envisioned. TFIIS could be only recruited to arrested Pol II complexes. Alternatively, it might associate and dissociate from the elongating Pol II, independently of transcription arrests, shifting to the cleavage-prone conformation of Pol II if need arises. The latter model would better account for the fact that GTP depletion by MPA, which is likely to promote arrest, does not strongly increase TFIIS occupancy.

The presence of TFIIS at nearly all Pol III-transcribed genes strongly suggests that it is a Pol III transcription factor. This hypothesis is supported by several lines of evidence. (1) TFIIS occupancy correlated well with that of Pol III genome-wide. (2) The occupancy profile of TFIIS closely followed that of Pol III on the *SCR1* gene, the longest class III gene. (3) TFIIS coimmunoprecipitated with Pol III after formaldehyde cross-linking. (4) A temperature-sensitive Pol II mutation (*rpb1-1*) strongly

reduced TFIIS enrichment at Pol II-transcribed genes but had no effect on its association with Pol III-transcribed genes. Conversely, a temperature-sensitive Pol III mutation (*rpc25-S100P*) diminished TFIIS binding at Pol III-transcribed genes with no effect on its association with Pol II-transcribed genes. Thus, the presence of TFIIS on Pol III-transcribed genes depends on Pol III activity and is independent of Pol II. (5) Under low-temperature conditions, the *dst1*-Δ mutation affected growth, and diminished Pol III and TFIIIB association with class III genes. (6) The *dst1* deletion impaired the translational suppression of *ade2-1* (a nonsense UAA mutant) by the $SUP4^{ochre}$ suppressor, and reduced the Pol III transcript levels under low-temperature conditions. (7) TFIIS improved Pol III transcription start site selection in vitro. Altogether, these data indicate that TFIIS is a bona fide component of the Pol III transcription machinery. It has been suggested previously that TFIIS could also play a role in Pol I transcription (Sawadogo et al. 1980a; Sawadogo et al. 1981; Schnapp et al. 1996), but a recent report demonstrated that the intrinsic cleavage activity of Pol I requires the Rpa12 subunit, sharing sequence homology with TFIIS (Kuhn et al. 2007). Our ChIP results showed that TFIIS was enriched on the Pol I-transcribed rDNA templates. TFIIS could thus be implicated in Pol I transcription. Interestingly, the Pol III-transcribed 5S rRNA genes, arranged in tandem with the 35S rRNA genes, were not enriched by TFIIS. The reason for the absence of TFIIS on 5S rRNA genes is unknown but could stem from the intragenic binding of TFIIIA factor that is required for 5S transcription.

As a Pol II elongation factor, TFIIS strictly depends on the RSADE domain, since inactivating this domain (or deleting the entire TFIIS) makes cells sensitive to NTP-depleting drugs (mycophenolate, 6-azauracil), which is generally seen as a consequence of a defective Pol II-associated cleavage (Exinger and Lacroute 1992; Ubukata et al. 2003). As a factor involved in Pol II initiation, TFIIS is needed for the full recruitment of Pol II to several promoters, especially in the absence of the Med31 subunit of Mediator. This, however, does not depends on the RSADE motif but is impaired in *dst1-R200A*, a mutation of the TFIIS Pol II-interacting domain (Guglielmi et al. 2007). Turning now to the Pol III-associated role(s) of TFIIS, we found that the RSADE motif is clearly important in this context, since the corresponding mutant form of TFIIS alters start site selection by Pol III in vitro, and since Pol III occupancy is strongly diminished in *dst1-E291A* cells grown at 16°C, with a strongly perturbed Pol III transcriptome. The reason why class III genes, the transcription of which is most diminished, differ in *dst1*-Δ and *dst1-E291A* is presently unknown but might be the consequence of the strong perturbation of Pol III transcription in the latter background. An indirect effect of altered Pol I or Pol II transcription is unlikely because *dst1* mutations do not affect 35S rRNA precursor transcription at 16°C (data not shown) and because Pol II occupancy defect is not increased in *dst1-E291A* compared with *dst1*-Δ. In contrast, *dst1-R200A* had a limited but significant effect on Pol II occupancy, with no effect at all on Pol III, suggesting that this mutation might specifically affect a Pol II-associated function (Guglielmi et al. 2007).

In the course of this study, we examined the possible implication of TFIIS in the different steps of in vitro Pol III transcription. Omitting TFIIS did not detectably influence Pol III elongation or cleavage activity in vitro but altered start site selection, suggesting that TFIIS might primarily act at the level of Pol III recruitment and/or transcription initiation. This is further supported by our ChIP assays showing that *dst1*-Δ strongly reduces the presence of Bdp1, the TFIIB-related subunit of the TFIIIB initiation factor of Pol III, with no effect on the TFIIIC initiation factor. In Pol III transcription, start site selection and initiation require the precise targeting of the enzyme by its initiation factors (TFIIIC and TFIIIB) and the opening of the transcription bubble around the start site (Kassavetis and Geiduschek 2006). TFIIS might participate in the initial steps of the Pol III transcription cycle in different ways. TFIIS could bend DNA and facilitate appropriate DNA binding by the transcription machinery as do Nhp6A and Nhp6B (Kassavetis and Steiner 2006). Alternatively, TFIIS could facilitate productive initiation by influencing the interaction of Pol III with basal factors for more accurate enzyme positioning. We favor the second hypothesis since direct binding of TFIIS to DNA has never been shown and since TFIIS and Pol III coimmunoprecipitated in cross-linked extracts.

One could be initially surprised to find TFIIS associated with class III genes, since Pol III has an intrinsic transcript cleavage activity that depends on Rpc11 subunit (Chedin et al. 1998; Alic et al. 2007). Rpc11 has a C-terminal Zn loop that bears an RSADE motif, critical for transcript cleavage, that closely resembles the C-terminal domain of TFIIS (Chedin et al. 1998). Our results suggest that both Rpc11 subunit and TFIIS are required for efficient Pol III transcription, but that their roles are not identical. Recombinant TFIIS (Fig. 7), in contrast to recombinant Rpc11 alone (Chedin et al. 1998), could not restore cleavage activity of Pol IIIΔ variant lacking Rpc11, suggesting that TFIIS does not participate in this reaction. Further, we found that Rpc11 and TFIIS bound throughout *SCR1* gene as Rpc160 (used as a proxy for Pol III) did. Moreover, TFIIS and Rpc11 coimmunoprecipitated in cross-linked extracts, as did Rpc160, indicating that their binding to class III genes is not mutually exclusive.

In conclusion, there is now mounting evidence that TFIIS controls several levels of DNA transcription. At a subset of Pol II-transcribed gene promoters, it could be recruited and act together with Mediator to recruit Pol II (Prather et al. 2005; Guglielmi et al. 2007; Kim et al. 2007), independently of its transcript cleavage stimulatory activity. TFIIS could also dynamically associate and dissociate from Pol II and stimulate the enzyme intrinsic RNA cleavage activity when needed. In addition, the present study shows that TFIIS is a Pol III transcription factor that stimulates Pol III transcription and may contribute to precise start site selection.

Materials and methods

Protein purification

DB3.1 *E. coli* cells containing either pDEST17-*DST1* or pDEST17-*dst1E291A* were grown at 30°C to 0.6 OD_{600}. Expression of the 6xHis-TFIIS fusion protein was induced by addition of 1 mM isopropyl-1-thio-β-D-galactopyranoside. Cells were harvested after 3 h of induction and resuspended in 20 mM HEPES buffer (pH 7.5) containing 10 µM $ZnCl_2$, 300 mM $NaCl_2$, 10% glycerol, 10 mM β-mercaptoethanol, and a set of protease inhibitors (phenyl-methyl-sulfonyl floride, Complete [Roche]). After lysis by sonication at 4°C, the lysate was clarified by centrifugation in a Beckman JA20 rotor for 25 min at 12,000 rpm. The 6xHis-TFIIS proteins were purified using an ÄKTA purifier (Amersham Biosciences) on a Hi Trap Chelating HP 5-mL column with a gradient of imidazole from 10 mM to 1 M. SDS-PAGE analysis, followed by Coomassie brilliant blue staining, showed that the fusion proteins were purified to near homogeneity.

ChIP and genome-wide ChIP–chip

Cross-linked chromatin was prepared essentially as described previously (Kuras and Struhl 1999; Kuras et al. 2003). Cells were grown exponentially to 0.6 OD_{600} and cross-linked with 1% formaldehyde for 10 min. The 3HA- and 13Myc-tagged proteins were immunoprecipitated with 12CA5 and 9E10 antibodies, respectively; Pol II was immunoprecipitated with 8WG16 anti-CTD antibody (Covance), and bound to IgG magnetic beads (Dynabead). Immune complexes were washed as described previously (Kuras and Struhl 1999). Cross-link reversal and DNA purification were performed as described (Kuras et al. 2003), except that the final elution was in 50 µL. Immunoprecipitated DNA was analyzed by quantitative real-time PCR on an ABI Prism 7000 or 7300 machine (Applied Biosystems). The PCR reactions were carried out in 25 µL containing 0.4 µM each primer, and 12.5 µL of mastermix SYBR green PCR reaction (Applied Biosystems). Relative quantification using a standard curve method was performed and the occupancy level for a specific fragment was defined as the ratio of immunoprecipitated DNA over total DNA. *GAL1* ORF region was used as a non-transcribed control.

Ligation-mediated PCR was done as described previously (Ren et al. 2000), except that amino-allyl conjugated dUTP (150 µM final) was used and only 30 cycles of PCR were performed. The PCR products were purified using a Microcon YM-30 filter. Amino-allyl modified DNA was recovered with 20 µL of H_2O and the DNA was lyophilized. DNA was labeled as previously described (Harismendy et al. 2003), except that unincorporated dyes were removed using a QIAquick PCR Purification Kit (Qiagen). Labeled DNA was recovered with 100 µL of buffer EB, and ethanol precipitated. Hybridization and washing conditions were as described previously (Lee et al. 2006). Microarrays were obtained from Agilent Technologies and feature 41,418 (G4486A) or 41,776 (G4493A) 60-mer oligonucleotide probe spots, with an average density of one probe each 266 bp of the yeast genome. Images of Cy5 and Cy3 fluorescence intensities were generated using the Genepix 4000B scanner (Molecular Devices) and extracted using GenePix Pro 6 software (Molecular Devices). At least two biological replicates were performed for each experiment.

Data analysis

The computational data analysis was performed in R using the *limma* package (Smyth et al. 2005) from the bioconductor project (http://www.bioconductor.org). After subtraction of the local background, the data from both channels were median normalized, and log ratios of signal intensities were generated for each feature. The log ratios were processed by fitting a linear model for each feature in order to calculate the average log ratio between replicates. *P*-values were then calculated by performing an empirical Bayes moderated *t*-statistic test, and adjusted for multiple testing by Benjamini and Hochberg false discovery rate (FDR) method. The complete raw data set is available at Array Express (http://www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-10. Visualization of ChIP-enriched genomic regions was performed using an adaptation of the Ringo package (Toedling et al. 2007). GO analysis was performed using GOstats package.

Pol III transcriptome analysis

Microarray hybridization was performed as described previously (Ciesla et al. 2007). Briefly, 20 µg of total RNA was reverse transcribed with specific primers designed to hybridize to the 3′ end of all mature tRNAs. cDNA was labeled with Cy3 and Cy5 dyes (Amersham) and hybridized on a Pol III-specific microarray (O. Harismendy, pers. comm.) harboring all the different tRNA genes, as well as all the other genes transcribed by Pol III.

In vitro transcription and primer extension assays

Standard Pol III in vitro transcription reactions on the *SUP4* template were performed as described previously (Ducrot et al. 2006; Alic et al. 2007) in 40 µL of transcription buffer. pRS316-*SUP4* plasmid (150 ng) was incubated for 20 min at 25°C in the presence of 100 ng of rTFIIIC, 20 ng of rTBP, 10 ng of rBdp1, 10 ng of rBrf1, 100 ng of highly purified Pol III, and 0.5 µg of purified B″ fraction or different concentrations of TFIIS protein (wild-type or E291A mutant). Transcription was started by the addition of 600 µM A/C/GTP, 300 µM UTP, 25 µM [α-^{32}P]UTP (400 Ci/mmol), and allowed to proceed for 15 min.

Primer extension assays were performed as described previously (Andrau and Werner 2001) on the *SUP4* template with 200 µM unlabeled ATP, CTP, UTP, and GTP. Single-round assays were performed as described, except that 125 µM unlabeled ATP, CTP, UTP, GTP, and 0.3 mg/mL heparin were used (Kassavetis and Steiner 2006).

The reaction products were analyzed by electrophoresis on denaturing polyacrylamide gels (7% for in vitro transcription and primer extension assay). Gels were autoradiographed using MR film with an intensifying screen (Kodak). Quantifications were performed with Image Quant software (Molecular Dynamics).

Acknowledgments

We thank M. Wery, D. Després, E. Shematorova, M. Boguta, K. Struhl, D. Stillman, and O. Gadal for yeast strains or plasmids; C. Kane for anti-TFIIS antibody; the SPI (CEA/Saclay) for monoclonal antibodies; C. Conesa, O. Harismendy, and S.G.F. (CEA/Evry) for Pol III transcriptome arrays and protocols; M. Riva for the yeast Pol II preparation and advice on Pol II transcription assays; C. Carles, N. Ayoub, and N. Alic for the yeast Pol IIIΔ preparation and advice in cleavage assays; and C. Ducrot for Pol III transcription factor preparations. We also thank O. Lefebvre and A. Briand for useful advice, and A. Sentenac and C. Mann for critical reading of the manuscript. Y.G. and M.M. were supported by a "Contrat de Formation par la Recherche" from the CEA.

References

Alic, N., Ayoub, N., Landrieux, E., Favry, E., Baudouin-Cornu, P., Riva, M., and Carles, C. 2007. Selectivity and proofread-

ing both contribute significantly to the fidelity of RNA polymerase III transcription. *Proc. Natl. Acad. Sci.* **104:** 10400–10405.

Andrau, J.C. and Werner, M. 2001. B″-associated factor(s) involved in RNA polymerase III preinitiation complex formation and start-site selection. *Eur. J. Biochem.* **268:** 5167–5175.

Awrey, D.E., Weilbaecher, R.G., Hemming, S.A., Orlicky, S.M., Kane, C.M., and Edwards, A.M. 1997. Transcription elongation through DNA arrest sites. A multistep process involving both RNA polymerase II subunit RPB9 and TFIIS. *J. Biol. Chem.* **272:** 14747–14754.

Awrey, D.E., Shimasaki, N., Koth, C., Weilbaecher, R., Olmsted, V., Kazanis, S., Shan, X., Arellano, J., Arrowsmith, C.H., Kane, C.M., et al. 1998. Yeast transcript elongation factor (TFIIS), structure and function. II: RNA polymerase binding, transcript cleavage, and read-through. *J. Biol. Chem.* **273:** 22595–22605.

Borukhov, S., Sagitov, V., and Goldfarb, A. 1993. Transcript cleavage factors from *E. coli*. *Cell* **72:** 459–466.

Braglia, P., Dugas, S.L., Donze, D., and Dieci, G. 2007. Requirement of Nhp6 proteins for transcription of a subset of tRNA genes and heterochromatin barrier function in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **27:** 1545–1557.

Chedin, S., Riva, M., Schultz, P., Sentenac, A., and Carles, C. 1998. The RNA cleavage activity of RNA polymerase III is mediated by an essential TFIIS-like subunit and is important for transcription termination. *Genes & Dev.* **12:** 3857–3871.

Chen, W., Tabor, S., and Struhl, K. 1987. Distinguishing between mechanisms of eukaryotic transcriptional activation with bacteriophage T7 RNA polymerase. *Cell* **50:** 1047–1055.

Ciesla, M., Towpik, J., Graczyk, D., Oficjalska-Pham, D., Harismendy, O., Suleau, A., Balicki, K., Conesa, C., Lefebvre, O., and Boguta, M. 2007. Maf1 is involved in coupling carbon metabolism to RNA Polymerase III transcription. *Mol. Cell. Biol.* **27:** 7693–7702.

Davis, C.A. and Ares Jr., M. 2006. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci.* **103:** 3262–3267.

Ducrot, C., Lefebvre, O., Landrieux, E., Guirouilh-Barbat, J., Sentenac, A., and Acker, J. 2006. Reconstitution of the yeast RNA polymerase III transcription system with all recombinant factors. *J. Biol. Chem.* **281:** 11685–11692.

Exinger, F. and Lacroute, F. 1992. 6-Azauracil inhibition of GTP biosynthesis in *Saccharomyces cerevisiae*. *Curr. Genet.* **22:** 9–11.

Fish, R.N. and Kane, C.M. 2002. Promoting elongation with transcript cleavage stimulatory factors. *Biochim. Biophys. Acta* **1577:** 287–307.

Geiduschek, E.P. and Kassavetis, G.A. 2001. The RNA polymerase III transcription apparatus. *J. Mol. Biol.* **310:** 1–26.

Guglielmi, B., Soutourina, J., Esnault, C., and Werner, M. 2007. TFIIS elongation factor and Mediator act in conjunction during transcription initiation *in vivo*. *Proc. Natl. Acad. Sci.* **104:** 16062–16067.

Harismendy, O., Gendrel, C.G., Soularue, P., Gidrol, X., Sentenac, A., Werner, M., and Lefebvre, O. 2003. Genome-wide location of yeast RNA polymerase III transcription machinery. *EMBO J.* **22:** 4738–4747.

Hausner, W., Lange, U., and Musfeldt, M. 2000. Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase. *J. Biol. Chem.* **275:** 12393–12399.

Kassavetis, G.A. and Geiduschek, E.P. 2006. Transcription factor TFIIIB and transcription by RNA polymerase III. *Biochem. Soc. Trans.* **34:** 1082–1087.

Kassavetis, G.A. and Steiner, D.F. 2006. Nhp6 is a transcriptional initiation fidelity factor for RNA polymerase III transcription *in vitro* and *in vivo*. *J. Biol. Chem.* **281:** 7445–7451.

Kettenberger, H., Armache, K.J., and Cramer, P. 2003. Architecture of the RNA polymerase II–TFIIS complex and implications for mRNA cleavage. *Cell* **114:** 347–357.

Kim, B., Nesvizhskii, A.I., Rani, P.G., Hahn, S., Aebersold, R., and Ranish, J.A. 2007. The transcription elongation factor TFIIS is a component of RNA polymerase II preinitiation complexes. *Proc. Natl. Acad. Sci.* **104:** 16068–16073.

Kuhn, C.D., Geiger, S.R., Baumli, S., Gartmann, M., Gerber, J., Jennebach, S., Mielke, T., Tschochner, H., Beckmann, R., and Cramer, P. 2007. Functional architecture of RNA polymerase I. *Cell* **131:** 1260–1272.

Kulish, D. and Struhl, K. 2001. TFIIS enhances transcriptional elongation through an artificial arrest site *in vivo*. *Mol. Cell. Biol.* **21:** 4162–4168.

Kuras, L. and Struhl, K. 1999. Binding of TBP to promoters *in vivo* is stimulated by activators and requires Pol II holoenzyme. *Nature* **399:** 609–613.

Kuras, L., Borggrefe, T., and Kornberg, R.D. 2003. Association of the Mediator complex with enhancers of active genes. *Proc. Natl. Acad. Sci.* **100:** 13887–13891.

Landrieux, E., Alic, N., Ducrot, C., Acker, J., Riva, M., and Carles, C. 2006. A subcomplex of RNA polymerase III subunits involved in transcription termination and reinitiation. *EMBO J.* **25:** 118–128.

Lee, T.I., Johnstone, S.E., and Young, R.A. 2006. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat. Protoc.* **1:** 729–748.

Malagon, F., Tong, A.H., Shafer, B.K., and Strathern, J.N. 2004. Genetic interactions of *DST1* in *Saccharomyces cerevisiae* suggest a role of TFIIS in the initiation–elongation transition. *Genetics* **166:** 1215–1227.

Martens, J.A., Laprade, L., and Winston, F. 2004. Intergenic transcription is required to repress the *Saccharomyces cerevisiae SER3* gene. *Nature* **429:** 571–574.

Mason, P.B. and Struhl, K. 2005. Distinction and relationship between elongation rate and processivity of RNA polymerase II *in vivo*. *Mol. Cell* **17:** 831–840.

Moqtaderi, Z. and Struhl, K. 2004. Genome-wide occupancy profile of the RNA polymerase III machinery in *Saccharomyces cerevisiae* reveals loci with incomplete transcription complexes. *Mol. Cell. Biol.* **24:** 4118–4127.

Nakanishi, T., Shimoaraiso, M., Kubo, T., and Natori, S. 1995. Structure–function relationship of yeast S-II in terms of stimulation of RNA polymerase II, arrest relief, and suppression of 6-azauracil sensitivity. *J. Biol. Chem.* **270:** 8991–8995.

Natori, S., Takeuchi, K., Takahashi, K., and Mizuno, D. 1973. DNA dependent RNA polymerase from Ehrlich ascites tumor cells. II. Factors stimulating the activity of RNA polymerase II. *J. Biochem.* **73:** 879–888.

Nonet, M., Scafe, C., Sexton, J., and Young, R. 1987. Eucaryotic RNA polymerase conditional mutant that rapidly ceases mRNA synthesis. *Mol. Cell. Biol.* **7:** 1602–1611.

Olmsted, V.K., Awrey, D.E., Koth, C., Shan, X., Morin, P.E., Kazanis, S., Edwards, A.M., and Arrowsmith, C.H. 1998. Yeast transcript elongation factor (TFIIS), structure and function. I: NMR structural analysis of the minimal transcriptionally active region. *J. Biol. Chem.* **273:** 22589–22594.

Opalka, N., Chlenov, M., Chacon, P., Rice, W.J., Wriggers, W., and Darst, S.A. 2003. Structure and function of the transcription elongation factor GreB bound to bacterial RNA polymerase. *Cell* **114:** 335–345.

Pokholok, D.K., Hannett, N.M., and Young, R.A. 2002. Exchange of RNA polymerase II initiation and elongation factors during gene expression *in vivo*. *Mol. Cell* **9:** 799–809.

Prather, D.M., Larschan, E., and Winston, F. 2005. Evidence that the elongation factor TFIIS plays a role in transcription initiation at *GAL1* in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **25:** 2650–2659.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309.

Roberts, D.N., Stewart, A.J., Huff, J.T., and Cairns, B.R. 2003. The RNA polymerase III transcriptome revealed by genome-wide localization and activity–occupancy relationships. *Proc. Natl. Acad. Sci.* **100:** 14695–14700.

Roberts, D.N., Wilson, B., Huff, J.T., Stewart, A.J., and Cairns, B.R. 2006. Dephosphorylation and genome-wide association of Maf1 with Pol III-transcribed genes during repression. *Mol. Cell* **22:** 633–644.

Sawadogo, M., Huet, J., and Fromageot, P. 1980a. Similar binding site for P37 factor on yeast RNA polymerases A and B. *Biochem. Biophys. Res. Commun.* **96:** 258–264.

Sawadogo, M., Sentenac, A., and Fromageot, P. 1980b. Interaction of a new polypeptide with yeast RNA polymerase B. *J. Biol. Chem.* **255:** 12–15.

Sawadogo, M., Lescure, B., Sentenac, A., and Fromageot, P. 1981. Native deoxyribonucleic acid transcription by yeast RNA polymerase–P37 complex. *Biochemistry* **20:** 3542–3547.

Schnapp, G., Graveley, B.R., and Grummt, I. 1996. TFIIS binds to mouse RNA polymerase I and stimulates transcript elongation and hydrolytic cleavage of nascent rRNA. *Mol. Gen. Genet.* **252:** 412–419.

Shilatifard, A., Conaway, R.C., and Conaway, J.W. 2003. The RNA polymerase II elongation complex. *Annu. Rev. Biochem.* **72:** 693–715.

Smyth, G.K., Michaud, J., and Scott, H.S. 2005. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21:** 2067–2075.

Toedling, J., Sklyar, O., and Huber, W. 2007. Ringo—An R/Bioconductor package for analyzing ChIP–chip readouts. *BMC Bioinformatics* **8:** 221; doi: 10.1186/1471-2105-8-221.

Ubukata, T., Shimizu, T., Adachi, N., Sekimizu, K., and Nakanishi, T. 2003. Cleavage, but not read-through, stimulation activity is responsible for three biologic functions of transcription elongation factor S-II. *J. Biol. Chem.* **278:** 8580–8585.

Wery, M., Shematorova, E., Van Driessche, B., Vandenhaute, J., Thuriaux, P., and Van Mullem, V. 2004. Members of the SAGA and Mediator complexes are partners of the transcription elongation factor TFIIS. *EMBO J.* **23:** 4232–4242.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B., et al. 2005. Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121:** 725–737.

Zaros, C. and Thuriaux, P. 2005. Rpc25, a conserved RNA polymerase III subunit, is critical for transcription initiation. *Mol. Microbiol.* **55:** 104–114.

Upon DNA damage, Sub1 regulates cell growth genes including RNA Polymerase III-transcribed genes

Audrey Suleau[1#], Arounie Tavenet[1#], Roberto Ferrari[2], Cécile Ducrot[1], Magali Michaut[1], Jean-Christophe Aude[1], Giorgio Dieci[2], Olivier Lefebvre[1], Christine Conesa[1]* and Joel Acker[1]**

[1]CEA, iBiTecS, Gif sur Yvette, F-91191, France.

[2]Dipartimento di Biochimica e Biologia Molecolare, Università degli Studi di Parma, 43100 Parma, Italy.

Contact

*Corresponding author:

Christine Conesa

Phone: + 33 1 69 08 37 96; Fax: +33 1 69 08 47 12

E-mail address: christine.conesa@cea.fr

**Corresponding author:

Joel Acker

Phone: + 33 1 69 08 65 61; Fax: +33 1 69 08 47 12

E-mail address: joel.acker@cea.fr

Additional footnotes

[#]These authors contributed equally to this work

Running Title: Sub1 regulates cell growth genes upon DNA damage

Abstract

Human PC4 and the yeast ortholog Sub1 have multiple functions in RNA polymerase (Pol) II transcription. Genome-wide mapping revealed that Sub1 is present on a subset of Pol II-transcribed genes implicated in protein synthesis as well as on Pol I and all Pol III-transcribed genes. Sub1 and PC4 are localized in both the nucleoplasm and the nucleolus. Upon DNA damage, Sub1 appears to act as a global regulator of genes involved in cellular growth, exerting both positive and negative effects on gene transcription. Sub1 is required to fully repress ribosomal protein and Pol III gene transcription. Using the Pol III transcription machinery as a model, we decipher the molecular mechanisms allowing Sub1 to stimulate the initiation and the reinitiation steps of the transcription cycle.

Introduction

PC4 plays an important role in various cellular processes, including transcription, DNA repair and replication (Ge and Roeder, 1994; Pan et al., 1996; Wang et al., 2004). First identified as a RNA polymerase (Pol) II co-activator (Ge and Roeder, 1994; Kretzschmar et al., 1994), PC4 was shown to interact with activators and components of the Pol II basal transcription machinery (Ge and Roeder, 1994; Malik et al., 1998) and to enhance activator-dependent transcription, stimulating both initiation and promoter escape (Fukuda et al., 2004). On the other hand, in the absence of TFIIH and TFIID, PC4 represses basal transcription (Malik et al., 1998; Werten et al., 1998). Sub1, the yeast ortholog of PC4, was characterized biochemically as a coactivator required for activated transcription in vitro (Henry et al. 1996) and genetically as a suppressor of certain TFIIB mutations (Knaus et al. 1996). Sub1 was found to interact directly with TFIIB and to inhibit the formation of TFIIB-TBP-DNA complex, suggesting a role in the release of TFIIB from the promoter (Knaus et al. 1996). Further investigations extended the role of PC4/Sub1 in transcription elongation and mRNA processing. Sub1 was shown to regulate enzymes modifying the CTD of the largest subunit of Pol II and might therefore enhance elongation (Calvo and Manley, 2005). Functional interactions between PC4/Sub1 and Cstf-64/Rna15 (Calvo and Manley, 2001) but also between Pta1 and Sub1 (He et al., 2003) established additional connections between Sub1/PC4 and mRNA processing. Recently, PC4 was found to be associated with chromatin and to be important for chromatin organization suggesting a more general role in transcription regulation (Das et al., 2006).

Apart from its role in transcription, PC4 was implicated in other cellular processes, through its capacity to bind tightly to melted DNA and to single-stranded DNA (ssDNA). A direct interaction between PC4 and XPG, a subunit of the nucleotide excision factor, was correlated with the genetic interaction between their yeast counterparts, Sub1 and Rad2, suggesting a role for PC4/Sub1 in the repair of oxydative DNA damage (Wang et al. 2004). Furthermore, PC4 can form complexes with human HSSB protein on ssDNA and influences its replication function in vitro (Pan et al., 1996).

In this study, we examined the general role of Sub1 in transcription. The genome-wide localization of Sub1 in exponentially growing yeast cells revealed its association to a subset of highly transcribed Pol II genes, to the rDNA gene and to all the genes transcribed by Pol III. We present evidence that Sub1 is involved in Pol III transcription initiation and re-

initiation processes in vitro. The results suggest that Sub1 functions as a global regulator of cell growth, upon DNA damage.

Results

Genome-wide analysis of Sub1 occupancy

A genome-wide location analysis by ChIP on chip of the gene occupancy by Sub1 was undertaken to define the gene targets of Sub1 in vivo. Chromatin immunoprecipitation (ChIP) assays were performed on epitope-tagged Sub1-3HA cross-linked chromatin from exponentially growing cells. Immunopurified DNA and DNA from whole-cell extracts were fluorescently labelled and competitively hybridized to DNA microarrays harbouring ORFs and intergenic regions. The ratio of fluorescence intensities at each site in the microarray provided a measure of the extent of Sub1 binding to a specific genomic locus. Data from three independent experiments were compiled. Many loci were found to be significantly enriched (991 loci with a p-value<0.01) in active growth conditions where Sub1 is not essential for cell viability. Approximately one-fourth of the enriched loci were located within ORF and the others corresponded to intergenic regions and to genes encoding non-translated RNAs. We first found that the *ACT1*, *PMA1*, *PYK1*, *ADH1* and snoRNA genes previously identified as DNA targets of Sub1 by ChIP and PCR amplification (Yang et al., 2005) were indeed enriched in our data. The distribution of the loci corresponding to snoRNA genes shown in Fig. 1A extended the binding of Sub1 to all the H/ACA box or C/D box snoRNA genes. We then used the GoTermFinder software to search for significantly enriched Gene Ontology (GO) terms associated with the enriched loci. Three over-represented GO categories (p-value<10^{-5}) were identified suggesting that Sub1 was preferentially bound to a subset of Pol II-transcribed genes encoding constituents of the cell wall, the nucleosome and the ribosome: Sub1 was found to be associated to 30 genes involved in cell wall structure (p-value<0.01) and to the genes encoding H2A, H2B, H3 and H4 histones (but not H1 histone). The third GO category (10% of the enriched loci) corresponded to genes encoding constituents of the ribosome, including *TEF1*, *TEF2* and 50 ribosomal protein (RP) genes (p-value<0.01). The distribution of the loci corresponding to RP genes is shown in Fig. 1A. Since RP and histone genes are among the most highly transcribed genes in exponentially growing cells, we determined the relationship between transcription rates and Sub1 occupancy. Interestingly, we found that most Sub1-associated Pol II genes were highly transcribed (Fig S1). Remarkably, these Pol II genes represented only two-third of the enriched loci. The other ones corresponded to the Loci Enriched by the Pol III Transcription Machinery (LEPTM) described previously in our laboratory using similar DNA microarrays (Harismendy et al., 2003). As shown in Fig. 1A, the LEPTM, comprising tRNA genes and other genes transcribed

by Pol III like 5S RNA gene, *SNR6*, *RPR1* and *SCR1* were significantly over-represented among the most enriched DNA regions, suggesting the association of Sub1 to all Pol III-transcribed genes in conditions of active growth.

To confirm the hypothesis of a role of Sub1 in regulating Pol III transcription in vivo, we performed ChIP assays to investigate the association of Sub1 with selected Pol III genes in exponentially growing cells. The DNA fragments immunopurified from epitope-tagged or wild-type Sub1 cross-linked chromatin were analyzed by PCR amplification. As shown in Fig. 1B, the ChIP of Sub1-3HA showed a clear enrichment for DNA fragments corresponding to the $tRNA^{Leu}$, 5S RNA, *SNR6* and *SCR1* genes transcribed by Pol III. As controls, we analyzed two other DNA fragments, a telomeric region of chromosome XV that gave background levels of amplification, and a Pol I promoter region, present in 100-150 copies in the yeast genome. Surprisingly, the rDNA fragment was clearly enriched in the ChIP of Sub1-3HA as compared to untagged Sub1. Note that in a previous work the same Pol I promoter DNA fragment gave background levels when the ChIP of Maf1 was analyzed by PCR amplification (Oficjalska-Pham et al., 2006). The arrays used had a poor coverage of the rDNA gene locus. Nevertheless, we found a significant enrichment of the different loci corresponding to that region suggesting that Sub1 associates at many locations throughout the rDNA gene. A conventional ChIP experiment on the rDNA locus (Fig. 1C) confirmed that Sub1 was enriched across all the rDNA region like Hmo1, a protein that plays a role in rRNA transcription (Hall et al., 2006). Altogether, the results suggested that Sub1 could be involved in the regulation of all three transcription systems.

Sub1 is localized in both the nucleoplasm and the nucleolus

In *S. cerevisiae*, 5S RNA genes are adjacent to each of the 100-150 rDNA transcription units and could therefore be nucleolar. Recent evidence also indicate that, though tRNA genes are dispersed in the genome, many of them are spatially clustered at or near the nucleolus, forming a subnuclear region specialized in Pol III transcription (Thompson et al., 2003). Because Sub1 was found to be present on both Pol I and Pol III-transcribed genes, we analyzed its intracellular localization using immunofluorescence. In exponentially growing cells, the fluorescence signal corresponding to Sub1 was uniformly distributed throughout the nucleus, overlapping the DAPI staining, and was also visualized in a region adjacent to the nucleoplasm that could delineate the nucleolus (Fig. 2A). We compared the localization of both Sub1 and Pol I. As expected, Pol I was excluded from the nucleoplasm, revealed by the DAPI staining (Fig. 2A). On the other hand, the signal corresponding to Sub1 overlapped

both the DAPI staining and the Pol I signals, showing that Sub1 was present in both the nucleoplasm and the nucleolus (Fig. 2A). Furthermore, in contrast to the clustered localization observed for tRNA genes (Thompson et al., 2003), Sub1 signals appeared to be uniformly distributed within the nucleolus. Little information is available about the cellular localization of components of the Pol III transcription machinery. We examined the cellular localization of Pol III and TFIIIC. In sharp contrast to Sub1, Pol III and TFIIIC were located essentially in the nucleoplasm (Fig. 2B and data not shown), suggesting that the presence of Sub1 in the nucleolus should not be attributed to its role in Pol III transcription.

Interestingly, two proteomic analyses undertaken to describe the protein content of human nucleoli identified PC4 as a nucleolar protein (Andersen et al., 2005; Coute et al., 2006). To corroborate these results, we carried out immunofluorescence confocal microscopy to compare the cellular localization of PC4 with that of Pol II or fibrillarin in HeLa cells. The cells were also labelled with the Syto RNASelect dye (invitrogen) that exhibits prominent nucleolar staining. The two optical sections, shown in Fig. S2, illustrate the co-localization of PC4 with Pol II within the nucleus and its partial co-localization with both fibrillarin and Syto RNASelect dye in some focal sections. We concluded that PC4 is mainly located in the nucleoplasm and is also present in a perinucleolar region.

Sub1-dependent genome-wide remodeling of gene expression under stress conditions

Little is known about the role of Sub1 on gene expression at the genome-wide level in yeast cells. Knocking down PC4 expression in HeLa cells by RNA interference has been recently shown to alter chromatin organization in vivo and to mostly up-regulate the expression of ≈180 genes (Das et al., 2006). Whereas there was no difference in the cell growth rate of *sub1Δ* cells in YPD at 30°C as compared to wild-type strain, a delay to reach the exponential phase was reproducibly observed for the *sub1Δ* strain when cells grown to exponential or stationary phase were diluted in a fresh medium (Fig. S3). Sub1 has also been shown to be required for resistance to oxidizing agents (Begley et al., 2002; Wang et al., 2004), suggesting that Sub1 may have some physiological importance under certain growth conditions. To confirm this hypothesis, wild-type and *sub1Δ* cells were plated at various dilutions on drug-supplemented YPD media and the cells were allowed to grow for several days at 30°C (data not shown). In our strain background, *sub1Δ* cells were indeed more sensitive than wild-type cells to hydrogen peroxide, but also to caffeine, camptothecin and to 4-nitroquinoline 1-oxide (4NQO, 0.2 mg/ml, a carcinogenic bulky alkylating agent). In contrast, the growth of *sub1Δ*

cells was not impaired in the presence of rapamycin (data not shown). Since Pol III transcription repression occurs in response to DNA damages (Ghavidel and Schultz, 2001) we focused our study on the modulation of gene expression upon 4NQO treatment. We used microarrays analysis to compare mRNA levels in exponentially growing wild-type and *sub1Δ* cells, after 1 h treatment with 4NQO. After spot quantification and normalization, gene expression levels were plotted as a function of Sub1 binding ratios as determined by our ChIP on chip experiments (Fig. 1A). As expected, only few genes were found to be differentially expressed when the two strains were exponentially grown in rich medium (Fig 3A left panel, 46 genes deregulated at least 4-fold with a p-value<0.01), in accordance with the fact that no cell growth defect could be detected in conditions of active growth when *SUB1* is deleted. In contrast, the distribution of gene expression ratios was clearly different upon 4NQO treatment showing that the transcriptional response to DNA damages was affected in the absence of Sub1. As shown in Fig. 3A (right panel), most of the differentially expressed genes were up-regulated in *sub1Δ* cells (215 out of the 261 genes deregulated at least 4-fold with a p-value<0.01), suggesting a global repression effect of Sub1. About half of them were bound by Sub1 demonstrating that Sub1 occupancy correlates with gene regulation. A GoTermFinder analysis revealed an enrichment in genes implicated in cell cycle process for the down-regulated genes, in keeping with the previously described involvement of Sub1 in DNA repair (Wang et al. 2004). The up-regulated genes were enriched in ribosome biogenesis and cell wall structure GO categories, confirming that most of the genes bound by Sub1 were up-regulated in *sub1Δ* cells upon 4NQO treatment. Note that several genes involved in protein biogenesis and chromatin-associated proteins were also found to be up-regulated upon knocking down PC4 expression in HeLa cells (Das et al., 2006). Sub1 could also regulate gene expression through indirect mechanisms. A third GO category corresponded to genes encoding proteins involved in amino acid metabolism whose expression is under the control of the Gcn4 transcriptional activator. This GO category could reflect an indirect effect since the *GCN4* gene is bound by Sub1, up-regulated in *sub1Δ* cells and induced at the translational level in response to DNA-damaging agents (Jelinsky et al., 2000).

It is known that RP genes are down-regulated in response to DNA-damaging agents (Jelinsky et al., 2000). As shown in Fig. 3A (right panel, black circles), the RP genes were significantly over-represented in the most up-regulated genes in response to 4NQO treatment. We thus wondered if the up-regulation of RP genes was due to a lesser repression of their expression upon 4NQO treatment in *sub1Δ* cells. A gene array displaying normalized ratios of each probe

according to the red-green colour-scale is shown in Fig. 3B (upper panel and supplementary Table S1). The global decrease of the levels of RP mRNAs observed upon 4NQO treatment in our wild-type strain (lane 1, median of the ratios of 0.49) did not occur in the *sub1Δ* strain (lane 2, median of the ratios of 1.1). This different efficiency in transcription repression upon 4NQO treatment led to higher levels of RP mRNAs in *sub1Δ* cells (lane 4, red, median of the ratios of 3.3) whereas there were no significant differences in RP gene expression between both strains in conditions of active growth (lane 3, median of the ratios of 0.9). These results suggested that the presence of Sub1 was required to fully repress RP gene expression in response to DNA damage.

To evaluate whether the presence of Sub1 was also important for Pol III transcription regulation in vivo, fluorescently labelled cDNAs from wild-type and *sub1Δ* strains were hybridized to a Pol III-specific microarray (Harismendy et al., manuscript in preparation) harbouring all the tRNA genes (one for each of the 52 different tRNA families described in Saccharomyces Genome Database) as well as all the other genes transcribed by Pol III. Normalization was performed with control probes (not shown) corresponding to three exogenous genes from *A. thaliana*. As expected, the levels of Pol III-synthesised RNAs globally decreased upon 4NQO treatment in the wild-type strain (Fig. 3B, lane 1, green, median of the ratios of 0.5, and supplementary Table S2). A less efficient repression of Pol III transcription occurred upon 4NQO treatment in the *sub1Δ* strain (lane 2, green, median of the ratios of 0.8) leading to higher levels of Pol III transcripts (lane 4, red, median of the ratios of 1.5) whereas there was no significant difference between both strains grown in exponential phase (lane 3, median of the ratios of 0.9). Even though Pol III-transcribed gene expression was modulated to a lesser extent than RP genes in response to 4NQO treatment, the presence of Sub1 was important to fully repress Pol III transcription in response to DNA damages.

Interestingly, although Sub1 was found to have a global negative effect on both RP and Pol III gene expression, this was not true for all of these genes. In several cases (Fig. 3B, lanes 3, green), lower levels of transcripts were observed in the absence of Sub1 in exponentially growing cells, suggesting a more complex role in vivo, with both a positive and negative effect on gene transcription.

Whether the presence of Sub1 was important for Pol I transcription regulation in vivo was investigated by primer extension on the 35S RNA precursor. As expected, in response to DNA damages, the transcription by Pol I was down-regulated upon 4NQO treatment, but we did not detect any significant difference in *sub1Δ* cells relative to wild-type cells (data not shown).

The above results indicated that Sub1 mediates the transcriptional regulation of a subset of genes involved in cellular growth, including the Pol III-transcribed genes.

Sub1/PC4 stimulates a yeast basal Pol III transcription system

The association of Sub1 to all Pol III-transcribed genes in vivo and the old observation of a stimulatory effect of PC4 on a human Pol III transcription system in vitro (Wang and Roeder, 1998) prompted us to test the effect of Sub1 in vitro using a basal transcription system. This system reconstituted with all recombinant factors and highly purified Pol III directs a low level of RNA synthesis and needs to be supplemented with partially purified fractions like B" (Kassavetis et al., 1992) or TFIIIE (Dieci et al., 1993; Ruth et al., 1996) to restore strong transcription rates (Ducrot et al., 2006). This suggested the existence of positive auxiliary factors enhancing Pol III transcription to maintain the high rates of Pol III transcription that are necessary to sustain rapid cell growth and proliferation. Western blot analysis revealed the presence of Sub1 in the B" fraction but not in the most purified TFIIIE fractions (Fig. S4A and data not shown) showing that Sub1 could not account for TFIIIE activity. As shown in Fig. 4A, increasing amounts of highly purified recombinant Sub1 (Fig. S4B), but not of a control protein, strongly stimulated the basal transcription of the *SUP4* $tRNA^{Tyr}$ (up to 4-fold) and 5S RNA (up to 6-fold) genes. Only specific Pol III transcription was stimulated since the presence of Sub1 did not change the non-specific transcription activity of Pol III on a poly d(A-T) template (data not shown). A strong stimulation (up to 5-fold) of the basal transcription of the *SUP4* $tRNA^{Tyr}$ gene was also observed with purified recombinant human PC4 (Fig. S4C). Note that a strong stimulation of Pol III transcription in vitro by Sub1 or PC4 was observed with all the yeast Pol III-transcribed genes that we have tested.

Sub1 is involved in the stimulation of Pol III transcription initiation and reinitiation in vitro

We examined the effect of Sub1/PC4 in transcription initiation by analyzing single round transcription on a *SUP4* tDNA template under the conditions that yield a 17-mer RNA. As previously described (Andrau and Werner, 2001), several transcription products corresponding to different start sites could be detected using recombinant rTFIIIB (Fig. 4B). In the presence of B" fraction, the transcriptional initiation specificity was restored, only the 17-mer RNA was observed and the transcription level was stimulated up to 3 times (Fig. 4B). In the presence of Sub1 or PC4, a 3-fold stimulation of the transcription was also observed but neither Sub1 nor PC4 had any effect on start site selection. We concluded that Sub1/PC4

directly affect the initiation efficiency of Pol III transcription (including pre-initiation complex assembly, promoter opening and promoter escape) whereas this effect was not observed in the human Pol III transcription system (Wang and Roeder, 1998).

PC4 was found to enhance and extend the interactions of TFIIIC with downstream promoter and termination DNA sequences (Wang and Roeder, 1998). We explored a possible role of Sub1 in the assembly of pre-initiation complex using gel shift assays. Sub1 enhanced the binding of rTFIIIC on the *SUP4* $tRNA^{Tyr}$ gene (up to 2.5 fold) and slightly decreased the electrophoretic mobility of TFIIIC-DNA complexes (Fig. 5A), suggesting the stabilization of TFIIIC-DNA complexes by Sub1 assembly. Only a faint smear corresponding to minor non-specific protein-DNA complexes was observed with Sub1 alone (Fig. 5A). Using larger amounts of Sub1, TFIIIC-Sub1-DNA complexes and Sub1-DNA complexes could be detected simultaneously (Fig. 5B). The stability of preformed TFIIIC-DNA or TFIIIC-Sub1-DNA complexes was compared after incubation at increasing temperatures ranging from 25°C to 50°C (Fig. 5B). The presence of Sub1 did not influence the thermosensitivity of TFIIIC-DNA complexes. However, the progressive appearance of larger amounts of Sub1-DNA complexes correlating with the decrease in the slow-migrating form of TFIIIC-DNA complexes clearly demonstrated the presence of Sub1, assembled with TFIIIC on the DNA probe.

The second step in the assembly of pre-initiation complexes is the recruitment of TFIIIB by TFIIIC which results in the formation of DNA complexes of slower electrophoretic mobility, as shown in Fig. 5A. In the presence of Sub1, the formation of TFIIIB-TFIIIC-DNA complexes was enhanced (up to 2 fold) and the slower electrophoretic mobility of the resulting DNA complexes suggested that Sub1 was assembled in the complexes. To test whether Sub1 could directly influence the assembly of rTFIIIB alone into pre-initiation complexes, we took advantage of the $tRNA^{Ileu}$ (TAT) gene which can be transcribed in a TFIIIC-independent manner (Dieci et al., 2000) and bound by rTFIIIB in gel shift assays (Fig 5C). The addition of Sub1 strongly stimulated the binding of TFIIIB to DNA resulting in the formation of a larger amount of slow-migrating complexes (Fig. 5C) clearly different from the non-specific protein-DNA complexes obtained with Sub1 alone on this particular DNA probe. The stimulation of TFIIIB assembly on $tDNA^{Ileu}$ by Sub1 was correlated with an increase in the transcription levels observed in the presence of Sub1 (up to 4-fold) using a minimal transcription system composed of rTFIIIB and Pol III (Fig. 4A, compare lanes 11 and 13). Furthermore, TFIIIC and Sub1 seemed to have additive stimulatory effects on transcription (Fig. 4A, lanes 11 to 14).

The strong stimulation of transcription obtained in the presence of Sub1 prompted us to examine whether Sub1 could promote Pol III transcription reinitiation. The transcription initiation frequency was determined for the *SUP4* $tRNA^{Tyr}$ gene (Fig. 4C) by comparing the output of multiple versus single round of transcription performed with limiting amounts of Pol III as previously described (Ferrari et al., 2004). The basal transcription system reconstituted with rTFIIIC could support only 1.9 cycles of transcription in 5 min (Fig. 4C). In contrast, more than 20 transcription cycles were obtained within 5 min in the presence of Sub1 or B" fraction (Fig. 4C). Similar results were obtained for the $tRNA^{Ileu}$ gene (ratio of 1.2 and 19, Fig. S4D), indicating that Sub1 played a role in the transcription reinitiation process.

Sub1 interacts with several components of the Pol III transcription system

Based on gel shift assays (Fig. 5), our results suggested that Sub1 helps TFIIIB and TFIIIC to assemble on tRNA genes, possibly through direct protein-protein interactions with both basal factors, at least in a DNA-dependent manner. To address this hypothesis, FarWestern experiments were performed as previously described (Chaussivert et al., 1995) using ^{35}S-labelled Sub1 (or PC4) as a probe. As expected since PC4 forms tightly associated homodimers (Werten and Moras, 2006), we found that Sub1 and PC4 interact with themselves and with each other (Fig. 6A and data not shown). Sub1 was also found to interact with Bdp1, a component of TFIIIB (Fig. 6B) and with τ138 and τ95, two subunits of TFIIIC (Fig. 6C), in good agreement with the co-immunopurification of PC4 with human TFIIIC (Wang and Roeder, 1998).

Discussion

In this work, we present evidence that Sub1 is involved in the regulation of a sizeable subset of highly transcribed genes mainly implicated in cell growth, including all the Pol III-transcribed genes. Sub1 is a stress response factor with a dual negative and positive role in transcription. Upon DNA damage, Sub1 was found to be required for repression of the genes implicated in protein synthesis. We show that Sub1 interacts with components of the Pol III transcription system and stimulates distinct steps of the transcription cycle in a reconstituted in vitro system.

Sub1 stimulates Pol III transcription through two mechanisms

Sub1/PC4, previously well established as a Pol II transcription co-activator (Ge and Roeder, 1994) strongly stimulates in vitro transcription by Pol III. This result confirms and extends an old observation of Roeder's laboratory on the stimulation of in vitro Pol III transcription by PC4 and Topo1, indicating that PC4 may stimulate the reinitiation process (Wang and Roeder, 1998). Using a transcription system reconstituted with all recombinant factors and highly purified Pol III, we clearly demonstrated that Sub1 (but not rTopo1, data not shown) is a reinitiation factor. Furthermore, we showed that Sub1 is also involved in the initiation step of the transcription cycle. Sub1 interacts directly with TFIIIB and TFIIIC assembly factors, forms ternary complexes with them and DNA, and stimulates factor binding to their cognate sites, resulting in an increased level of Pol III transcription. This outline corresponds to the recruitment model for gene activation that has been reported for many Pol II activators (Ptashne and Gann, 1997). Consistently, PC4 was reported to enhance in vitro the formation of the PIC on Pol II promoters resulting, but only in the presence of an activator, in transcription stimulation due to an increase of the initiation and promoter escape steps (Fukuda et al., 2004; Malik et al., 1998). Recent analysis of the co-crystal structure of PC4 with single-stranded DNA further showed that most of the bases point outward, allowing the DNA in PC4-DNA complexes to be much more accessible for sequence-specific recognition by other proteins (Werten and Moras, 2006), such as p53 on its specific DNA sequences (Banerjee et al., 2004). Sub1 directly activates Pol II (Henry et al., 1996) and Pol III transcription in vitro. Based on all these results, we predict that Sub1 may be a component of the Pol II preinitiation complex.

The 3-fold stimulation of Pol III transcription initiation was much lower than the overall effect of Sub1 observed on multiple versus single rounds of transcription suggesting a role of

Sub1/PC4 in another step of the transcription cycle. Consistently, Sub1 was found to relieve the Pol III reinitiation defect observed with recombinant factors and therefore can be considered as a Pol III and Pol II reinitiation factor. The properties of Sub1/PC4 perfectly fit with previous Pol III reinitiation models (Dieci and Sentenac, 2003). This led us to propose a novel reinitiation strategy (see Fig. 7) where the following properties of Sub1/PC4 have been taken into account: (1) Sub1 stabilizes the association of TFIIIB and TFIIIC on DNA, (2) Sub1 interacts directly with at least two components of the Pol III transcription machinery, (3) PC4 and TFIIIC facilitate termination and PC4 extends TFIIIC interactions with downstream DNA regions and termination sequences (Wang and Roeder, 1998), (4) PC4 is able to bend DNA (Batta and Kundu, 2007). All these properties could directly contribute to the PIC-assisted reinitiation strategy that involves the coupling of Pol III transcription termination and reinitiation (Dieci and Sentenac, 1996). Upon DNA bending, Sub1/PC4 could help Pol III to be directly transferred from the terminator to the promoter regions as it has been proposed in the facilitated recycling pathway (Dieci and Sentenac, 1996). One is inclined to extend these conclusions to Pol II transcription reinitiation that plays an important role in transcriptional regulation. The molecular mechanisms of Pol II reinitiation are much more complex than for Pol III transcription reinitiation because mRNA capping, splicing and polyadenylation occur co-transcriptionally and form part of the transcription cycle (Proudfoot et al., 2002). Furthermore, Pol II transcription termination is directly connected to the 3'-end RNA processing machinery (Proudfoot, 2004) that could facilitate efficient reinitiation of transcription (Orphanides and Reinberg, 2002). Interestingly, Sub1/PC4 is involved in several steps of the Pol II transcription cycle in accordance with a role in Pol II transcription reinitiation. For instance, Sub1 interacts directly with a component of the polyadenylation factor in a way that influences Pol II transcription termination (Calvo and Manley, 2001). We could also wonder whether the effect of Sub1 in PIC stabilisation (Henry et al., 1996; Malik et al., 1998) could not result in the activation of the PIC-assisted Pol II reinitiation process that has been described in yeast (Yudkovsky et al., 2000).

Sub1 is a transcriptional regulator of genes involved in cellular growth and DNA damage response

Genome-wide occupancy studies revealed that the gene targets of Sub1 are essentially dedicated to translation, chromatin structure and RNA modifications. We propose to consider Sub1 as a global regulator of highly transcribed genes mainly involved in cellular growth, and as a stress response factor that exerts a complex role in transcription.

The present study revealed that Sub1 is not dedicated to one particular transcription system. Sub1 was detected on all the genes transcribed by Pol I and III as well as on a subset of Pol II-transcribed genes, encoding ribosomal proteins, histones and snoRNA. As described recently (Abruzzi et al., 2004), we could not rule out the fact that a part of the signal detected in ChIP assays could be due to a possible interaction of Sub1 with RNA. However, Sub1 binds directly to DNA, Sub1 is mainly present on intergenic regions, it interacts physically and functionally with Pol II and Pol III transcription initiation machineries and Sub1 occupancy correlates with gene regulation. All these results indicate that Sub1 is associated directly with its DNA targets in vivo, whose expression results in increased capacity for protein synthesis, a fundamental determinant of cell growth and proliferation. Although postulated, there is no clear evidence indicating whether the balanced supply of the ribosomal components involves a coordinated regulation of the three transcription machineries. The association of Sub1 with the rDNA locus, the RP genes and the 5S DNA makes Sub1 a promising candidate to play a role in such a coordinated regulation, even if the role of Sub1 on Pol I transcription is still speculative and needs to be further investigated. Sub1 is clearly present in the nucleolus and this cellular localization could not be attributed to its role in Pol III transcription. PC4 is also present in a subnuclear region close to the nucleolus. However, there is no evidence that Sub1 plays a direct role in Pol I transcription. We do not favor the hypothesis that Sub1 was associated with the rDNA transcription units because of the presence of Pol II-transcribed genes on this locus. Sub1 was enriched all along the rDNA region and it was found to interact directly (our unpublished data) with one of the specific subunits of Pol I and possibly with H3 histone, a component of the UAF transcription factor.

Second, we showed that Sub1 occupancy on both Pol III and Pol II-transcribed genes correlates with differential gene expression at least under stress conditions. In this work, we demonstrated that the presence of Sub1 is important for cell survival and transcription regulation in vivo in response to DNA damage induced by the UV-mimetic agent 4NQO. In that case, Sub1 appears to act as a stress response factor. On the other hand, it has been reported that the expression of Sub1 mRNA rapidly increases after the inoculation of quiescent yeast cells in rich medium (Radonjic et al., 2005). In keeping with this observation, we repeatedly observed that *sub1Δ* cells need a much extended lag phase before resuming normal growth. It would be interesting to analyze the possible role of Sub1 during this lag phase or during recovery from many poor growth conditions that were found to repress transcription (nutrients or serum starvation, mitotic repression, secretory pathway defects,

oxydative stress, DNA damages, chemical treatments with diverse drugs). Further investigations should be performed to reveal other growth conditions that may mobilize Sub1 for cellular survival.

Last, our results confirmed that like the NC2 cofactor (Willy et al., 2000) or Mot1 (Dasgupta et al., 2002), Sub1/PC4 plays a dual negative and positive complex role in transcription. We demonstrated that upon DNA damage Sub1 is required for full repression of RP and Pol III-transcribed genes, whereas knocking down PC4 expression using siRNA revealed mainly a negative role on gene expression (Das et al., 2006). Consistently, PC4 was shown to repress Pol II basal transcription in vitro (Malik et al., 1998). On the other hand, a positive role of Sub1/PC4 on gene expression has also been reported in vivo (Knaus et al., 1996; Marroquin et al., 2005) and Sub1 (this work) and PC4 (Ge and Roeder, 1994) are able to stimulate in vitro transcription systems reconstituted on naked DNA. We have to keep in mind that Sub1/PC4 is a multifunctional protein that plays important roles in diverse cellular processes. PC4 is involved in chromatin condensation (Das et al., 2006) that interferes with transcription. Further investigations in a chromatin context may help to understand the molecular mechanisms of Sub1/PC4 in transcription regulation.

Materials and Methods

Yeast Strains

All strains used in this study are described in the Supplemental Data available with this article on line.

DNA binding and in vitro transcription assays

Transcriptions were performed as described in Ducrot et al. (2006) with 10 ng of rTFIIIC, 20 ng of rTBP, 10 ng of rBrf1, 10 ng of rBdp1 or 0.5μg of partially purified B" fraction, 100 ng of highly purified Pol III and 40 ng of rTFIIIA when 5S template was used. Briefly, rTFIIIC, rTFIIIB and Pol III were incubated for 20 min at 25°C, in the presence or not of the indicated amounts of rSub1 or rPC4. Transcription was then allowed to proceed for 10 min at 25°C by the addition of NTPs and transcripts were analyzed by electrophoresis on a 8% polyacrylamide gel, 8M urea. Facilitated transcription reinitiation was performed as described (Ferrari et al., 2004). Heparin at a 200 μg/ml concentration was used in single-round transcription reactions. Protein-DNA interactions were monitored by gel shift assays as described previously (Ducrot et al., 2006) using a ^{32}P-labelled DNA fragment carrying the tRNA^{Ileu} or the *SUP4* tRNA^{Tyr} genes as probes. Transcripts or DNA complexes were visualized with a Typhoon 9200 Imager (Amersham Biosciences).

Chromatin immunoprecipitation, microarray hybridization and data analysis.

Chromatin immunoprecipitation and PCR were performed as previously described (Harismendy et al., 2003) using a wild-type strain or a strain expressing a 3-HA tagged version of Sub1. The oligonucleotides used for DNA amplification are described in our Supplemental Data. For ChIP on chip experiments, DNA from wild-type or Sub1-3HA strains were competitively hybridized on DNA microarrays as described (Harismendy et al., 2003). Data from three independent experiments were compiled. Gene expression was monitored on wild-type or *sub1Δ* cells grown to exponential phase or incubated in the presence of 4NQO (1 μg/μl) for 1h. RNA extraction and microarray hybridization were performed as previously described (Conesa et al., 2005). Data analysis are described in the Supplemental Data. The complete raw data set and the analyzed data are available at http://www.ncbi.nlm.nih.gov/geo (GEO accession: GSE11124 and GSE11054).

Acknowledgements:

We are grateful to A. Sentenac for helpful discussions and for improving the manuscript, A. Peyroche and B. Le Tallec for helpful suggestions. We thank G. Le Roux and R. Courbereytte for technical assistance and advices. We thank M. Teichmann for the generous gift of purified PC4 and the pet-PC4 expression plasmid, M. Riva for providing anti A190 antibodies, Franck Amiot, Peggy Maltere, Amélie Robert for providing the DNA chip (CEA, IRCM). This work was funded by grant ANR-07-BLAN-0039-01 from the French National Research Agency and by grant 3982 from the Association pour la Recherche contre le Cancer. A.T. was supported by the International PhD Program of the CEA.

References

Andersen, J.S., Lam, Y.W., Leung, A.K., Ong, S.E., Lyon, C.E., Lamond, A.I. and Mann, M. (2005) Nucleolar proteome dynamics. *Nature*, **433**, 77-83.

Andrau, J.C. and Werner, M. (2001) B"-associated factor(s) involved in RNA polymerase III preinitiation complex formation and start-site selection. *Eur J Biochem*, **268**, 5167-5175.

Banerjee, S., Kumar, B.R. and Kundu, T.K. (2004) General transcriptional coactivator PC4 activates p53 function. *Mol Cell Biol*, **24**, 2052-2062.

Batta, K. and Kundu, T.K. (2007) Activation of p53 function by Human Transcriptional Coactivator PC4: Role of Protein-Protein Interaction, DNA Bending and Posttranslational modifications. *Mol Cell Biol*.

Begley, T.J., Rosenbach, A.S., Ideker, T. and Samson, L.D. (2002) Damage recovery pathways in Saccharomyces cerevisiae revealed by genomic phenotyping and interactome mapping. *Mol Cancer Res*, **1**, 103-112.

Calvo, O. and Manley, J.L. (2001) Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination. *Mol Cell*, **7**, 1013-1023.

Calvo, O. and Manley, J.L. (2005) The transcriptional coactivator PC4/Sub1 has multiple functions in RNA polymerase II transcription. *Embo J*, **24**, 1009-1020.

Chaussivert, N., Conesa, C., Shaaban, S. and Sentenac, A. (1995) Complex interactions between yeast TFIIIB and TFIIIC. *J Biol Chem*, **270**, 15353-15358.

Conesa, C., Ruotolo, R., Soularue, P., Simms, T.A., Donze, D., Sentenac, A. and Dieci, G. (2005) Modulation of yeast genome expression in response to defective RNA polymerase III-dependent transcription. *Mol Cell Biol*, **25**, 8631-8642.

Coute, Y., Burgess, J.A., Diaz, J.J., Chichester, C., Lisacek, F., Greco, A. and Sanchez, J.C. (2006) Deciphering the human nucleolar proteome. *Mass Spectrom Rev*, **25**, 215-234.

Das, C., Hizume, K., Batta, K., Kumar, B.R., Gadad, S.S., Ganguly, S., Lorain, S., Verreault, A., Sadhale, P.P., Takeyasu, K. and Kundu, T.K. (2006) Transcriptional coactivator PC4, a chromatin-associated protein, induces chromatin condensation. *Mol Cell Biol*, **26**, 8303-8315.

Dasgupta, A., Darst, R.P., Martin, K.J., Afshari, C.A. and Auble, D.T. (2002) Mot1 activates and represses transcription by direct, ATPase-dependent mechanisms. *Proc Natl Acad Sci U S A*, **99**, 2666-2671.

Dieci, G., Duimio, L., Coda-Zabetta, F., Sprague, K.U. and Ottonello, S. (1993) A novel RNA polymerase III transcription factor fraction that is not required for template commitment. *J Biol Chem*, **268**, 11199-11207.

Dieci, G., Percudani, R., Giuliodori, S., Bottarelli, L. and Ottonello, S. (2000) TFIIIC-independent in vitro transcription of yeast tRNA genes. *J Mol Biol*, **299**, 601-613.

Dieci, G. and Sentenac, A. (1996) Facilitated recycling pathway for RNA polymerase III. *Cell*, **84**, 245-252.

Dieci, G. and Sentenac, A. (2003) Detours and shortcuts to transcription reinitiation. *Trends Biochem Sci*, **28**, 202-209.

Ducrot, C., Lefebvre, O., Landrieux, E., Guirouilh-Barbat, J., Sentenac, A. and Acker, J. (2006) Reconstitution of the yeast RNA polymerase III transcription system with all recombinant factors. *J Biol Chem*, **281**, 11685-11692.

Ferrari, R., Rivetti, C., Acker, J. and Dieci, G. (2004) Distinct roles of transcription factors TFIIIB and TFIIIC in RNA polymerase III transcription reinitiation. *Proc Natl Acad Sci U S A*, **101**, 13442-13447.

Fukuda, A., Nakadai, T., Shimada, M., Tsukui, T., Matsumoto, M., Nogi, Y., Meisterernst, M. and Hisatake, K. (2004) Transcriptional coactivator PC4 stimulates promoter escape and facilitates transcriptional synergy by GAL4-VP16. *Mol Cell Biol*, **24**, 6525-6535.

Ge, H. and Roeder, R.G. (1994) Purification, cloning, and characterization of a human coactivator, PC4, that mediates transcriptional activation of class II genes. *Cell*, **78**, 513-523.

Geiduschek, E.P. and Kassavetis, G.A. (2001) The RNA polymerase III transcription apparatus. *J Mol Biol*, **310**, 1-26.

Ghavidel, A. and Schultz, M.C. (2001) TATA binding protein-associated CK2 transduces DNA damage signals to the RNA polymerase III transcriptional machinery. *Cell*, **106**, 575-584.

Hall, D.B., Wade, J.T. and Struhl, K. (2006) An HMG protein, Hmo1, associates with promoters of many ribosomal protein genes and throughout the rRNA gene locus in Saccharomyces cerevisiae. *Mol Cell Biol*, **26**, 3672-3679.

Harismendy, O., Gendrel, C.G., Soularue, P., Gidrol, X., Sentenac, A., Werner, M. and Lefebvre, O. (2003) Genome-wide location of yeast RNA polymerase III transcription machinery. *Embo J*, **22**, 4738-4747.

He, X., Khan, A.U., Cheng, H., Pappas, D.L., Jr., Hampsey, M. and Moore, C.L. (2003) Functional interactions between the transcription and mRNA 3' end processing machineries mediated by Ssu72 and Sub1. *Genes Dev*, **17**, 1030-1042.

Henry, N.L., Bushnell, D.A. and Kornberg, R.D. (1996) A yeast transcriptional stimulatory protein similar to human PC4. *J Biol Chem*, **271**, 21842-21847.

Jelinsky, S.A., Estep, P., Church, G.M. and Samson, L.D. (2000) Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes. *Mol Cell Biol*, **20**, 8157-8167.

Kassavetis, G.A., Joazeiro, C.A., Pisano, M., Geiduschek, E.P., Colbert, T., Hahn, S. and Blanco, J.A. (1992) The role of the TATA-binding protein in the assembly and function of the multisubunit yeast RNA polymerase III transcription factor, TFIIIB. *Cell*, **71**, 1055-1064.

Knaus, R., Pollock, R. and Guarente, L. (1996) Yeast SUB1 is a suppressor of TFIIB mutations and has homology to the human co-activator PC4. *Embo J*, **15**, 1933-1940.

Kretzschmar, M., Kaiser, K., Lottspeich, F. and Meisterernst, M. (1994) A novel mediator of class II gene transcription with homology to viral immediate-early transcriptional regulators. *Cell*, **78**, 525-534.

Malik, S., Guermah, M. and Roeder, R.G. (1998) A dynamic model for PC4 coactivator function in RNA polymerase II transcription. *Proc Natl Acad Sci U S A*, **95**, 2192-2197.

Marroquin, C.E., Wai, P.Y., Kuo, P.C. and Guo, H. (2005) Redox-mediated upregulation of hepatocyte iNOS transcription requires coactivator PC4. *Surgery*, **138**, 93-99.

Oficjalska-Pham, D., Harismendy, O., Smagowicz, W.J., Gonzalez de Peredo, A., Boguta, M., Sentenac, A. and Lefebvre, O. (2006) General repression of RNA polymerase III transcription is triggered by protein phosphatase type 2A-mediated dephosphorylation of Maf1. *Mol Cell*, **22**, 623-632.

Orphanides, G. and Reinberg, D. (2002) A unified theory of gene expression. *Cell*, **108**, 439-451.

Pan, Z.Q., Ge, H., Amin, A.A. and Hurwitz, J. (1996) Transcription-positive cofactor 4 forms complexes with HSSB (RPA) on single-stranded DNA and influences HSSB-dependent enzymatic synthesis of simian virus 40 DNA. *J Biol Chem*, **271**, 22111-22116.

Proudfoot, N. (2004) New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr Opin Cell Biol*, **16**, 272-278.
Proudfoot, N.J., Furger, A. and Dye, M.J. (2002) Integrating mRNA processing with transcription. *Cell*, **108**, 501-512.
Ptashne, M. and Gann, A. (1997) Transcriptional activation by recruitment. *Nature*, **386**, 569-577.
Radonjic, M., Andrau, J.C., Lijnzaad, P., Kemmeren, P., Kockelkorn, T.T., van Leenen, D., van Berkum, N.L. and Holstege, F.C. (2005) Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon S. cerevisiae stationary phase exit. *Mol Cell*, **18**, 171-183.
Ruth, J., Conesa, C., Dieci, G., Lefebvre, O., Dusterhoft, A., Ottonello, S. and Sentenac, A. (1996) A suppressor of mutations in the class III transcription system encodes a component of yeast TFIIIB. *Embo J*, **15**, 1941-1949.
Thompson, M., Haeusler, R.A., Good, P.D. and Engelke, D.R. (2003) Nucleolar clustering of dispersed tRNA genes. *Science*, **302**, 1399-1401.
Wang, J.Y., Sarker, A.H., Cooper, P.K. and Volkert, M.R. (2004) The single-strand DNA binding activity of human PC4 prevents mutagenesis and killing by oxidative DNA damage. *Mol Cell Biol*, **24**, 6084-6093.
Wang, Z. and Roeder, R.G. (1998) DNA topoisomerase I and PC4 can interact with human TFIIIC to promote both accurate termination and transcription reinitiation by RNA polymerase III. *Mol Cell*, **1**, 749-757.
Werten, S. and Moras, D. (2006) A global transcription cofactor bound to juxtaposed strands of unwound DNA. *Nat Struct Mol Biol*, **13**, 181-182.
Werten, S., Stelzer, G., Goppelt, A., Langen, F.M., Gros, P., Timmers, H.T., Van der Vliet, P.C. and Meisterernst, M. (1998) Interaction of PC4 with melted DNA inhibits transcription. *Embo J*, **17**, 5103-5111.
Willy, P.J., Kobayashi, R. and Kadonaga, J.T. (2000) A basal transcription factor that activates or represses transcription. *Science*, **290**, 982-985.
Yang, P.K., Hoareau, C., Froment, C., Monsarrat, B., Henry, Y. and Chanfreau, G. (2005) Cotranscriptional recruitment of the pseudouridylsynthetase Cbf5p and of the RNA binding protein Naf1p during H/ACA snoRNP assembly. *Mol Cell Biol*, **25**, 3295-3304.
Yudkovsky, N., Ranish, J.A. and Hahn, S. (2000) A transcription reinitiation intermediate that is stabilized by activator. *Nature*, **408**, 225-229.

Figure legends

Figure 1: Sub1 is present on Pol I, Pol II and Pol III-transcribed genes. (A) Genome wide binding of Sub1. Immunopurified DNA and DNA from Sub1-3HA whole-cell extracts, fluorescently labelled, were competitively hybridized to DNA microarrays. The distribution of medium ranks of Cy3/Cy5 fluorescence ratios is represented as histograms. Loci corresponding to LEPTM, RP genes and snoRNA genes are represented as indicated (black curves) within the global distribution of loci (grey curve). The full-scale shows the distribution of all the loci. (B) ChIP analysis. Cross-linked chromatin (Input) from wild-type (WT) or Sub1-3HA exponentially grown cells were immunoprecipitated with antibodies specific to the epitope (IP). DNA enrichment of the indicated genes was assessed by PCR as described in Supplemental data. (C) Sub1 binds to the rDNA locus. Immunoprecipitation from wild-type or Sub1-13myc cross-linked chromatin was analyzed via quantitative PCR. The amounts of immunoprecipitated DNA expressed as a value relative to that of the Input are shown as histograms. Sequence elements within the rDNA unit (Non-Transcribed Spacer 1 and 2, the 5S RNA and 35S RNA transcripts) and the positions of the DNA fragments amplified to analyze the immunoprecipitates are schematically represented.

Figure 2: Sub1 but not Pol III colocalizes with Pol I in the nucleolus. Sub1 and Pol I (A) or Pol III and Pol I (B) were localized by immunofluorescence in exponentially growing cells. The cells were analyzed by phase contrast light microscopy (visible), by DNA staining (DAPI) and by immunofluorescence using antibodies specific to Sub1, Pol I or Pol III, as indicated. Overlays from the indicated combinations of images are shown.

Figure 3: Differential gene expression in *sub1*□ strain upon 4NQO treatment. Microarrays analysis was used to compare the RNA levels of essentially all Pol III- and Pol II transcribed genes in exponentially growing (-4NQO) wild-type (WT) and *sub1Δ* cells, or after 1 h of treatment with 4NQO (+4NQO). (A) Expression ratios of *sub1Δ*/WT were plotted as a function of Sub1 binding ratios. The distribution of RP genes is shown (black circle). (B) Expression ratios of RP genes or Pol III-transcribed genes in the indicated combinations of strains are presented according to the red-green color scale.

Figure 4: Sub1 is a Pol III transcription activator in vitro. (A) Sub1 stimulates the minimal Pol III transcription system. In vitro transcription of the *SUP4*-tDNA^{Tyr} gene (lanes 1

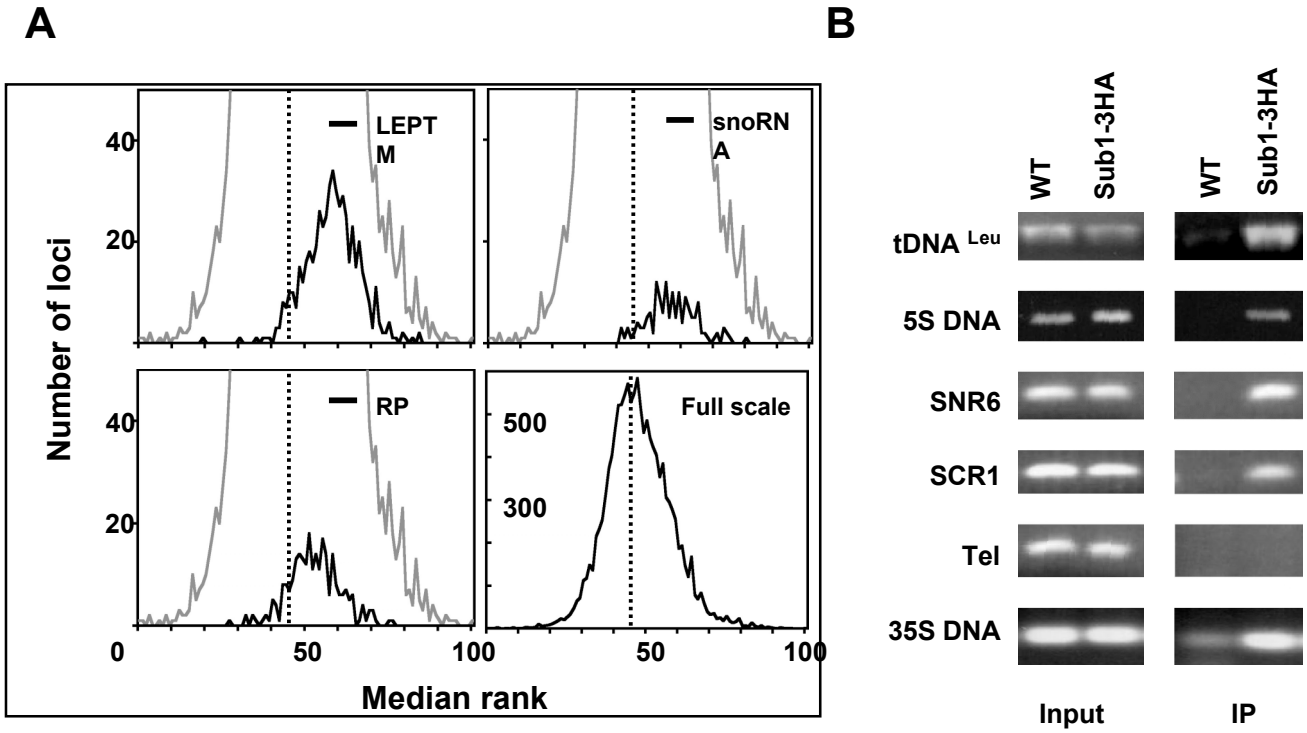
to 6), 5S RNA gene (lanes 7 to 10) and tDNA^{Ileu} gene containing a TATA-box (lanes 11 to 14) was carried out in the presence of rTBP, rBrf1, rBdp1, rTFIIIC (except for lanes 11 and 13), rTFIIIA (lanes 7 to 10 only), purified Pol III and varying amounts of purified rSub1 (lanes 2-5: 10 ng, 30ng, 80 ng, 240 ng; lanes 8-9: 40 ng, 150 ng; lanes 13-14: 100 ng) or purified rRsc4 (200 ng in lanes 6 and 10). (B) Sub1/PC4 stimulates transcription initiation. Stable pre-initiation complexes (PIC) were formed on the *SUP4*-tDNA^{Tyr} gene using rTFIIIC, rTBP, rBrf1 and purified Pol III in the presence of rBdp1 (lanes 1, 3 and 4) or of B" fraction (lane 2). rSub1 or rPC4 (respectively 100 and 60 ng) were added in lanes 3 and 4, respectively and transcriptions were carried out for 10 minutes in the absence of GTP to form a 17-mer ternary complex. 14/15-mer and 17-mer transcripts were quantified in each lane and ratios relative to lane 1 are indicated. (C) Sub1 enhances Pol III transcription reinitiation. PICs were assembled on the *SUP4*-tDNA^{Tyr} gene for 20 min in the presence of rTFIIIC and reconstituted TFIIIB with rTBP, rBrf1, and rBdp1 (lanes 1-4), or B" (lanes 5-6). rSub1 (100 ng) was added in lanes 3-4. Pol III (10 ng) was then added together with a mixture lacking CTP, and the incubation was continued for 20 min. Transcription was then resumed by the addition of CTP, either in the presence (+) or in the absence (-) of heparin, and the incubation was continued for 5 min. The ratios of multiple round (MR) versus single round (SR) of transcription are indicated.

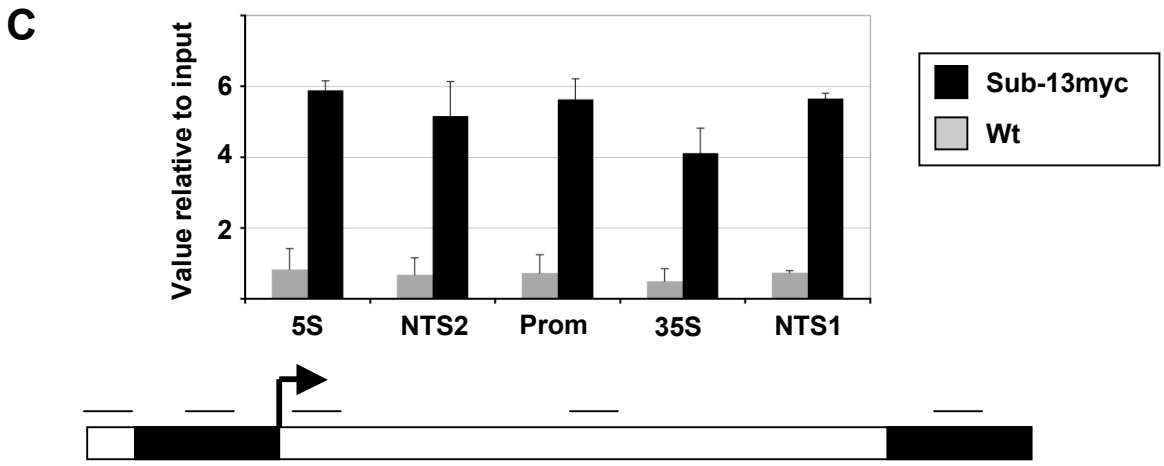
Figure 5: Sub1 enhances the assembly of TFIIIC-TFIIIB-tDNA complexes. (A) Sub1 helps TFIIIC to recruit TFIIIB on tDNA. rTFIIIC (C), rTFIIIB (B), rSub1 (S) were incubated as indicated with the *SUP4*-tDNA^{Tyr} probe. The positions of protein-tDNA complexes visualized by gel retardation assays are indicated. Lane 1: control without protein. (B) Sub1 does not affect TFIIIC-tDNA complexes stability. rTFIIIC was preincubated at 25°C with the *SUP4*-tDNA^{Tyr} probe alone (lanes 1-4) or in the presence of rSub1 (200 ng in lanes 6-11) and then further incubated for 10 minutes at the indicated temperatures. Lane 5: control without protein. (C) Sub1 stimulates the assembly of TFIIIB on tDNA. rTFIIIB and rSub1 were incubated as indicated with the tDNA^{Ileu} probe containing a TATA-box. Lane 1: control without protein.

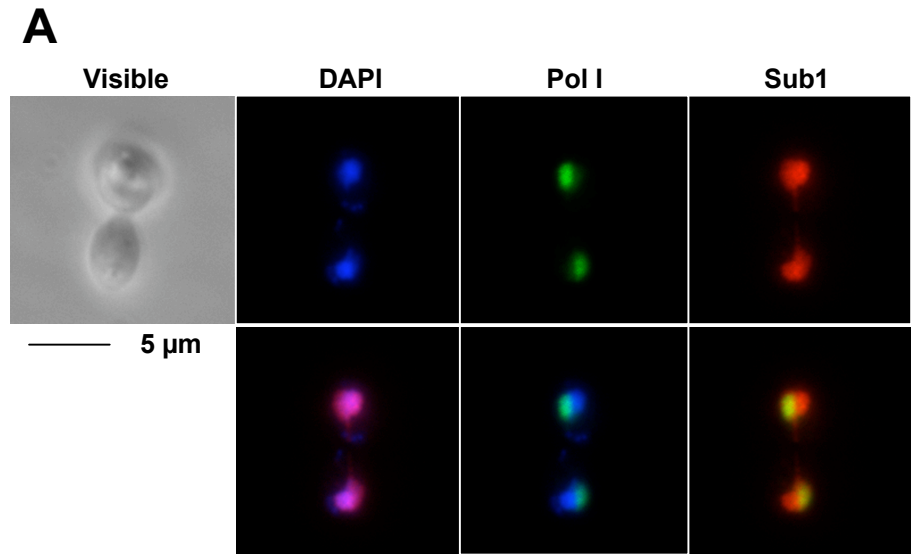
Figure 6: Sub1 interacts with components of the Pol III transcription machinery. rSub1 or rPC4 (A) and subunits of rTFIIIB (B) or of rTFIIIC (C) were subjected to SDS-PAGE, transferred onto a membrane, stained with Ponceau S (lanes P) and then probed with ^{35}S-labelled Sub1 (lanes ^{35}S). Labelled protein complexes were revealed by autoradiography.

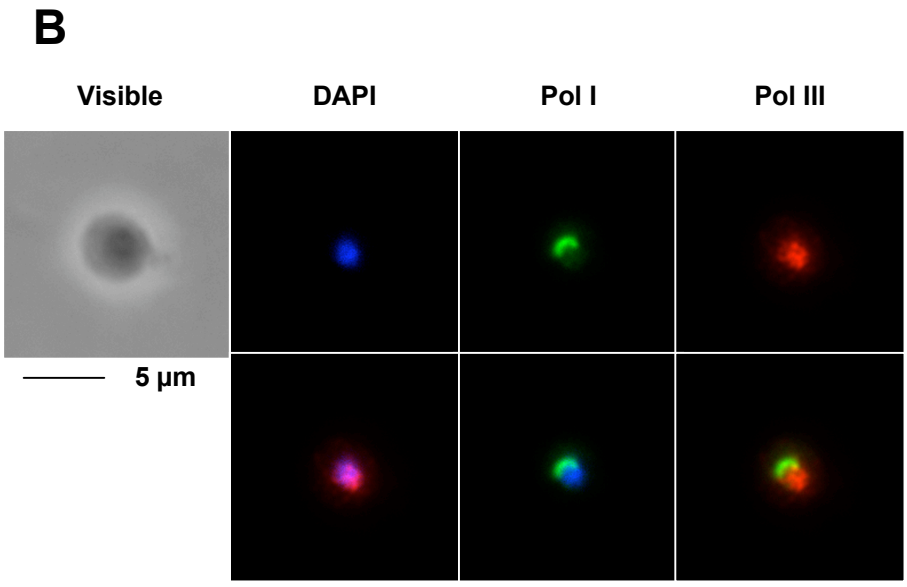
Figure 7: Reinitiation model in the RNA polymerase III transcription system.

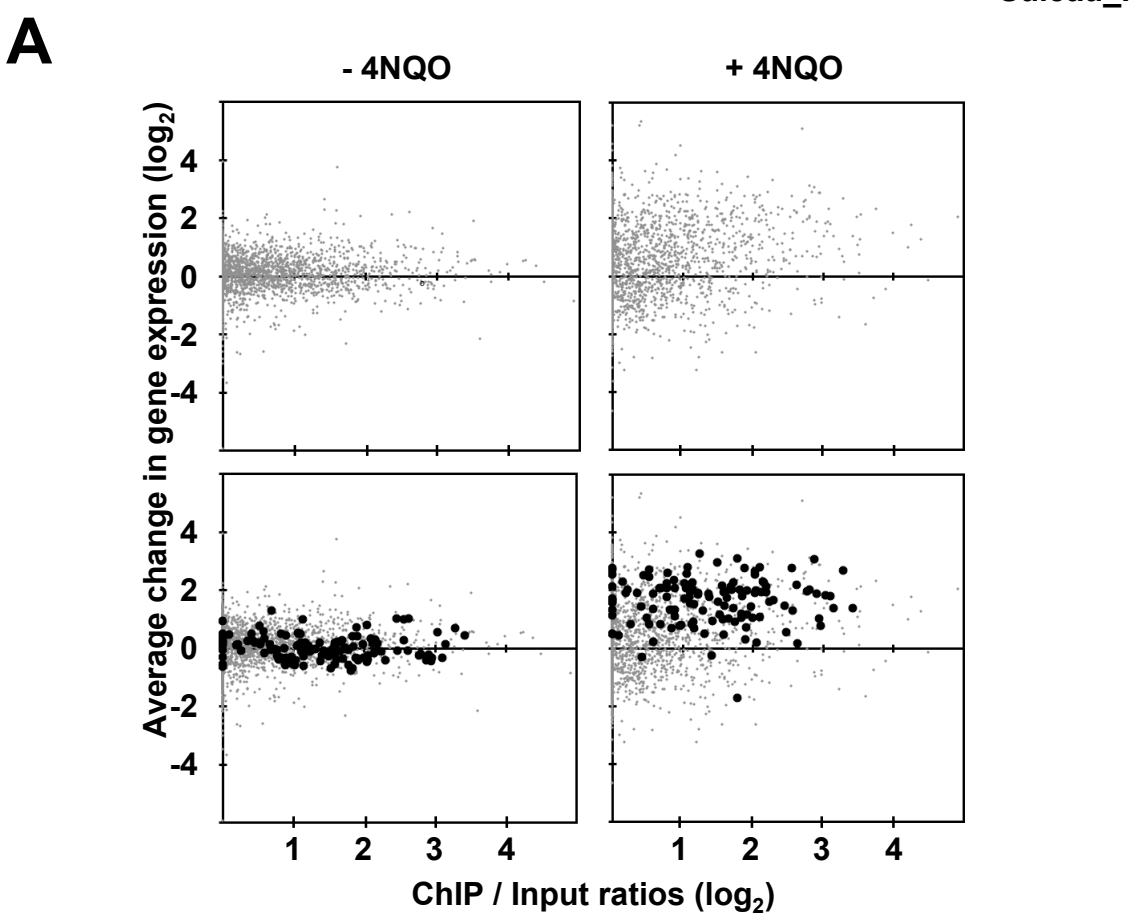
The assembly of a transcription complex on a yeast tRNA gene is depicted. The tDNA transcription unit and the components of the Pol III transcription system are schematically represented. TFIIIC binds to the A and B blocks, the intragenic promoter elements and directs the assembly of TFIIIB on tDNA. Binding of TFIIIB upstream of the transcription start site bends DNA, promotes Pol III recruitment resulting in the formation of the preinitiation complex (PIC) and in transcription initiation (Geiduschek and Kassavetis, 2001). The presence of Sub1, that may multimerize along the DNA (Werten and Moras, 2006), enhances PIC formation by DNA bending and through multiple interactions with t138 and t95 (TFIIIC), Bdp1 (TFIIIB) and possibly Rpc128 (Pol III, preliminary data). DNA bending mediated by Sub1 facilitates the direct transfer of Pol III from the terminator to the promoter region thus contributing to the facilitated recycling pathway (Dieci and Sentenac, 2003).


A
LEPTM
snoRNA
RP
Full scale
Number of loci
Median rank
B
WT
Sub1-3HA
tDNA Leu
5S DNA
SNR6
SCR1
Tel
35S DNA
Input
IP

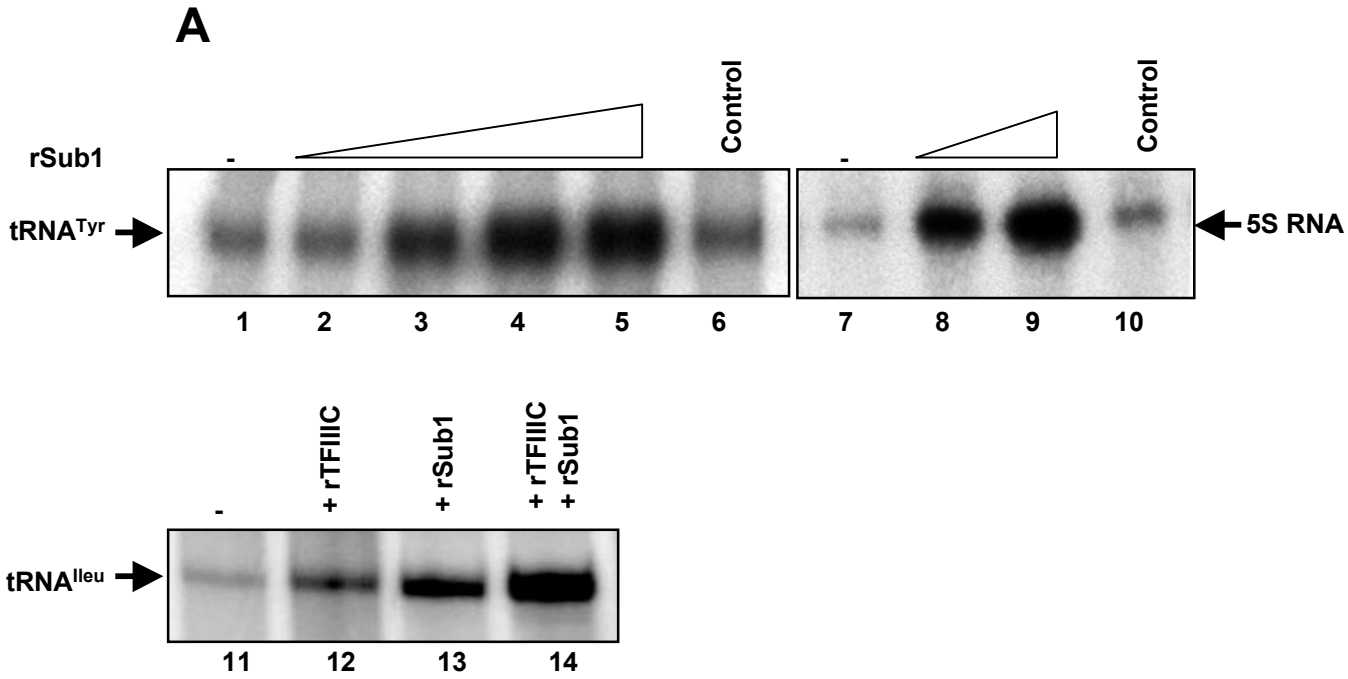

C
Value relative to input
Sub-13myc
Wt
5S
NTS2
Prom
35S
NTS1

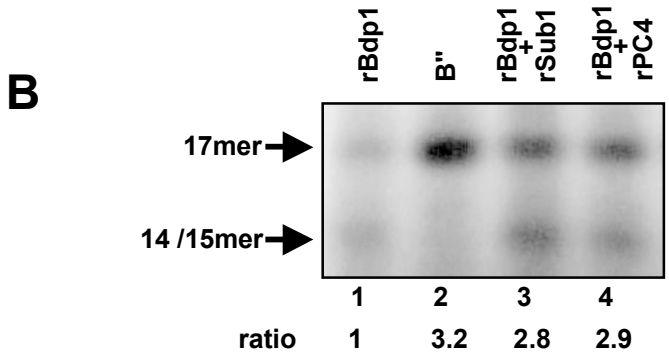

A
Visible
DAPI
Pol I
Sub1
5 µm

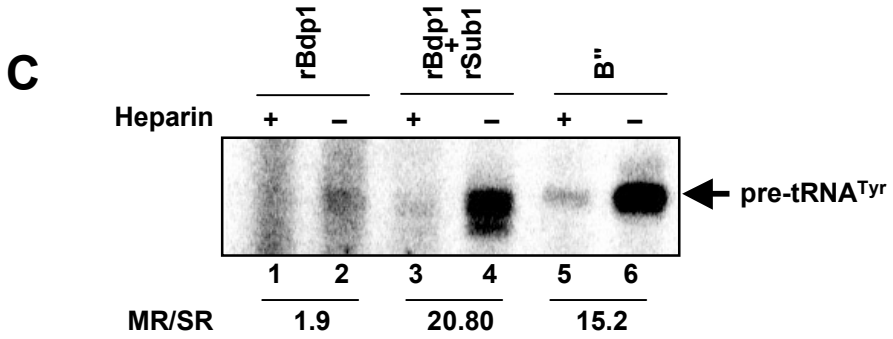

B
Visible
DAPI
Pol I
Pol III
5 µm

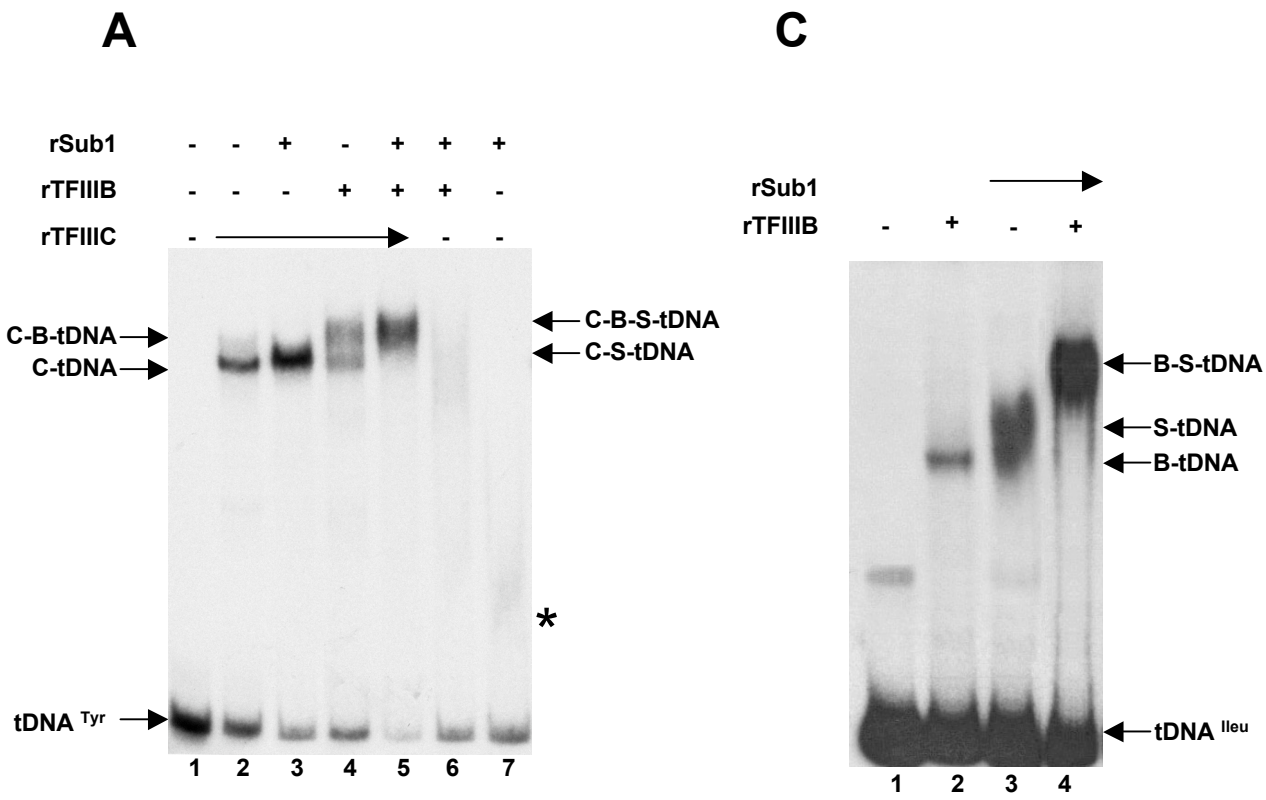

A
- 4NQO
+ 4NQO
Average change in gene expression (log2)
4
2
0
-2
-4
1
2
3
4
ChIP / Input ratios (log2)
B
RP genes
1
2
3
4
WT + 4NQO / WT - 4NQO
sub1Δ + 4NQO / sub1Δ - 4NQO
sub1Δ - 4NQO / WT - 4NQO
sub1Δ + 4NQO / WT + 4NQO
Pol III-transcribed genes
Fold change (log2)
3
2
1
0
-1
-2
-3
1
2
3
4
WT + 4NQO / WT - 4NQO
sub1Δ + 4NQO / sub1Δ - 4NQO
sub1Δ - 4NQO / WT - 4NQO
sub1Δ + 4NQO / WT + 4NQO

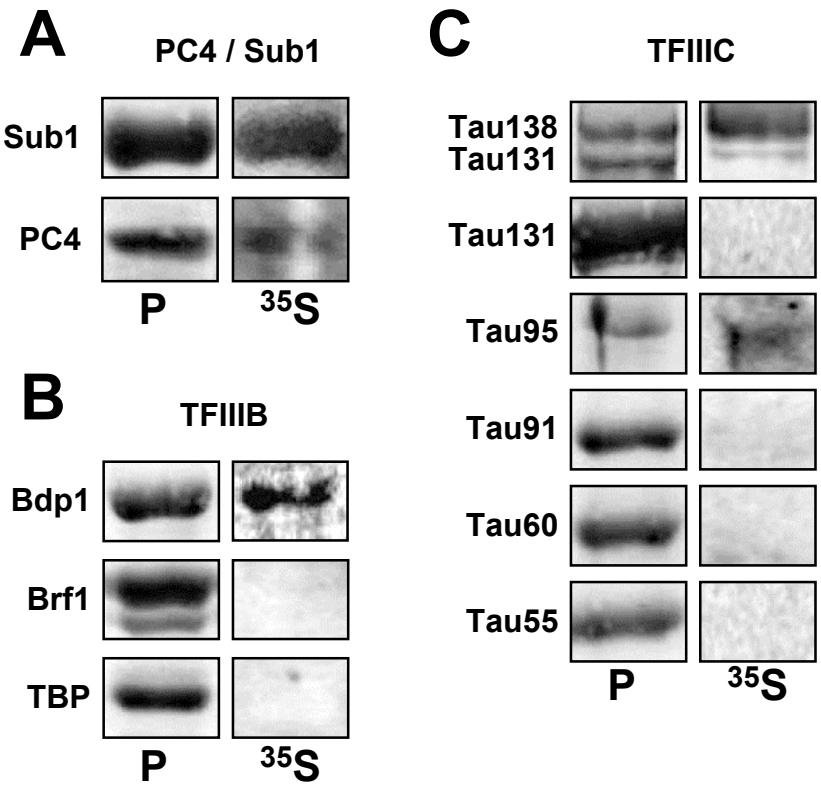
Suleau_Fig 4


A
rSub1
-
Control
-
Control
tRNATyr
5S RNA
1 2 3 4 5 6 7 8 9 10
-
+ rTFIIIC
+ rSub1
+ rTFIIIC + rSub1
tRNAIleu
11 12 13 14

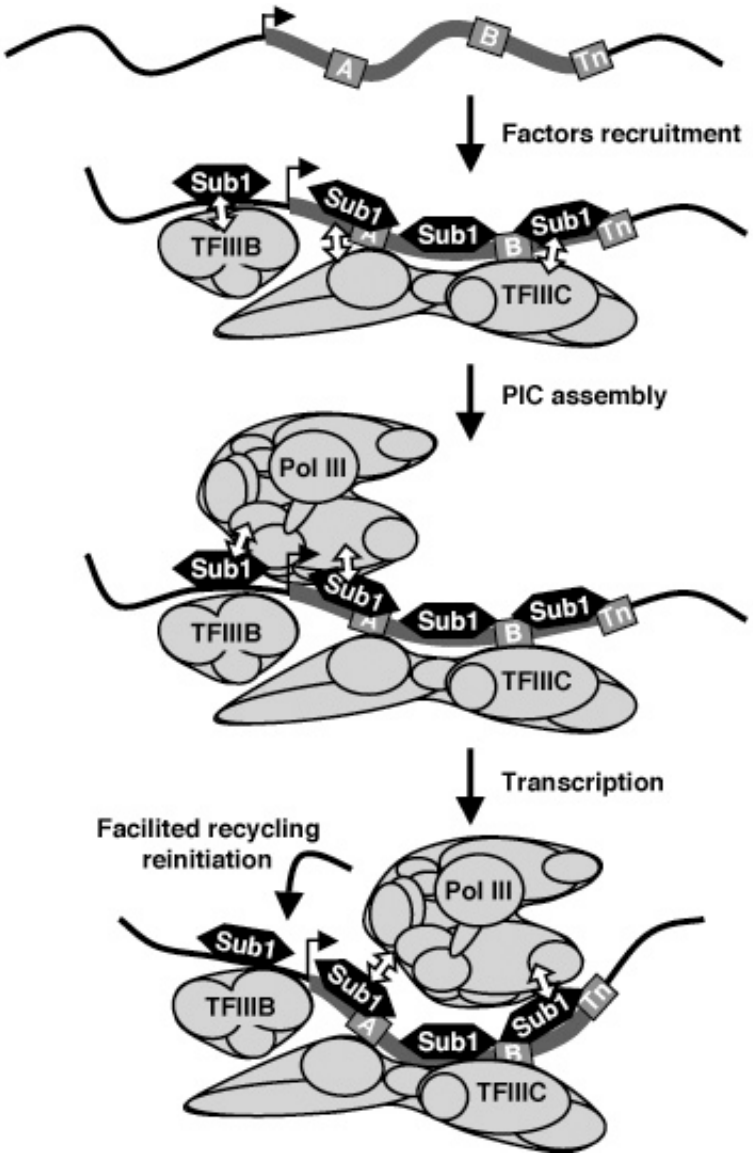

B
rBdp1
B"
rBdp1 + rSub1
rBdp1 + rPC4
17mer
14 /15mer
1 2 3 4
ratio 1 3.2 2.8 2.9


C
rBdp1
rBdp1 + rSub1
B"
Heparin
+ – + – + –
pre-tRNATyr
1 2 3 4 5 6
MR/SR
1.9
20.80
15.2


A
rSub1
rTFIIIB
rTFIIIC
C-B-tDNA
C-tDNA
tDNA Tyr
C-B-S-tDNA
C-S-tDNA
*
1 2 3 4 5 6 7
C
rSub1
rTFIIIB
B-S-tDNA
S-tDNA
B-tDNA
tDNA Ileu
1 2 3 4
B
rSub1
rTFIIIC
T (°C)
25 40 45 50 25 30 35 40 45 50
C-tDNA
S-tDNA
tDNATyr
1 2 3 4 5 6 7 8 9 10 11


A
PC4 / Sub1
Sub1
PC4
P
35S
B
TFIIIB
Bdp1
Brf1
TBP
P
35S
C
TFIIIC
Tau138
Tau131
Tau131
Tau95
Tau91
Tau60
Tau55
P
35S

Suleau_Fig 7


A
B
Tn
Factors recruitment
Sub1
TFIIIB
TFIIIC
PIC assembly
Pol III
Transcription
Facilited recycling
reinitiation

Analyse de données transcriptome et protéome pour l'étude des réponses aux stress oxydants et aux métaux lourds

Résumé :

Ce travail a pour objet l'analyse de données transcriptome et protéome pour l'étude des réponses aux stress oxydants et aux métaux lourds, en particulier chez la cyanobactérie *Synechocystis*. Cet organisme procaryote permet notamment d'aider à la compréhension des plantes tout en étant facilement manipulable génétiquement.

La démarche a d'abord consisté à analyser les réponses transcriptionnelles des gènes de *Synechocystis* en conditions de stress, notamment en présence de cadmium ou de peroxyde d'hydrogène. Des méthodes de prédiction d'interactions protéine-protéine ont ensuite été développées afin de construire un réseau d'interactions. Ce dernier a été comparé à un réseau d'interactions identifiées expérimentalement, notamment en termes de structure. Puis il a été complété avec les données de transcriptome précédemment analysées, afin d'obtenir une vision plus intégrée des différents phénomènes et d'étudier la dynamique des modules fonctionnels.

Les résultats font apparaître différentes phases dans les réponses transcriptionnelles, ainsi que des groupes fonctionnels de protéines en interaction et co-exprimées. De plus, l'automatisation d'une méthode de classification mixte hiérarchique-pyramidale est proposée. Une méthode d'identification de biais de composition entre des groupes de protéines a aussi été développée. Par ailleurs, un outil de prédiction d'interactions protéine-protéine, applicable à toutes les espèces séquencées, a été développé. Ce logiciel open-source, InteroPorc, présente l'avantage d'être flexible, puisqu'il peut s'appliquer à différents jeux d'interactions sources. En outre, l'outil est facilement utilisable en ligne à travers une interface web.

Mots clés :
bioinformatique, intégration, prédiction, transcriptome, protéome, interactions, réseau, logiciel

---

Transcriptomic and proteomic data analysis to study responses to oxidative stress and heavy metals

Summary :

This work aims at studying responses to oxidative stress and heavy metals through transcriptomic and proteomic data analysis, in particular in the cyanobacterium *Synechocystis*. This organism is a prokaryote largely studied which notably enables to improve the understanding of plants and is easy to manipulate genetically.

The approach first involved analysing the transcriptional responses of *Synechocystis*' genes in stress conditions, particularly in the presence of cadmium or hydrogen peroxide. Methods to predict protein-protein interactions were then developed in order to construct an interaction network. This network was compared to an experimental network in terms of structure. It was then complemented with transcriptomic data previously analysed in order to obtain a more integrated view of the different phenomena and to study the dynamics of functional modules.

The results show different phases in the transcriptional responses as well as functional groups of interacting and co-expressed proteins. In addition, the automation of a mixed hierarchical-pyramidal classification method is proposed. A method to identify composition biases between groups of proteins was also developed. Furthermore, a protein-protein interaction prediction tool was developed, of use for all sequenced species. This open-source software, InteroPorc, has been made available and has the great advantage of being flexible since it can be applied to different source interactions. Furthermore this tool can be easily run online through a web interface (http ://biodev.extra.cea.fr/interoporc/).

Keywords :
bioinformatics, integration, prediction, transcriptome, proteome, interactions, network, software